

# Entropy bounds on state estimation for stochastic non-linear systems under information constraints

Christoph Kawan and Serdar Yüksel\*

December 26, 2016

## Abstract

This paper studies state estimation over noisy channels for stochastic non-linear systems. We consider three estimation objectives, a strong and a weak form of almost sure stability of the estimation error as well as quadratic stability in expectation. For all three objectives, we derive lower bounds on the smallest channel capacity  $C_0$  above which the objective can be achieved with an arbitrarily small error. Lower bounds are obtained via a dynamical systems (through a novel construction of a dynamical system), an information-theoretic and a random dynamical systems approach. The first two approaches show that for a large class of systems, such as additive noise systems,  $C_0 = \infty$ , i.e., the estimation objectives cannot be achieved via channels of finite capacity. The random dynamical systems approach is shown to be operationally non-adequate for the problem, since it yields finite lower bounds  $C_0$  under mild assumptions. Finally, we prove that a memoryless noisy channel in general constitutes no obstruction to asymptotic almost sure state estimation with arbitrarily small errors, when there is no noise in the system.

**Keywords:** State estimation; non-linear systems; topological entropy; metric entropy; information theory; dynamical systems

**AMS Classification:** 93E10, 93E15, 93C10, 37A35

## 1 Introduction

State estimation over noisy channels is a first step towards a complete theory of control of non-linear systems over noisy channels. The results in this area have either almost exclusively considered linear systems, or in the non-linear case only been on deterministic systems over deterministic channels, with few exceptions.

State estimation over digital channels was studied in [28, 30] for linear discrete-time systems in a stochastic framework with the objective to bound the estimation error in probability. In these works, the inequality

$$C \geq H(A) := \sum_{\lambda \in \sigma(A)} \max\{0, n_\lambda \log |\lambda|\} \quad (1)$$

for the channel capacity  $C$  was obtained as a necessary and almost sufficient condition. Here  $A$  is the dynamical matrix of the system and the summation is over its eigenvalues  $\lambda$  with multiplicities  $n_\lambda$ .

Some relevant studies that have considered non-linear systems are the following. The papers [24], [35] and [29] studied state estimation for non-linear deterministic systems and noise-free channels. In [24], Liberzon and Mitra characterized the critical bit rate  $C_0$  for exponential state estimation with a given exponent  $\alpha \geq 0$  for a continuous-time system on a compact subset  $K$  of its state space. As a measure for

---

\*C. Kawan is with the Faculty of Computer Science and Mathematics, University of Passau, 94032 Passau, Germany (e-mail: christoph.kawan@uni-passau.de). S. Yüksel is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada, K7L 3N6 (e-mail: yuksel@mast.queensu.ca)

$C_0$ , they introduced a quantity named estimation entropy  $h_{\text{est}}(\alpha, K)$ , which coincides with the topological entropy on  $K$  when  $\alpha = 0$ , but for  $\alpha > 0$  is no longer a purely topological quantity. Furthermore, they derived an upper bound  $R$  of  $h_{\text{est}}(\alpha, K)$  in terms of  $\alpha$ , the dimension of the state space and a Lipschitz constant of the dynamical system. They also provided an algorithm accomplishing the estimation objective with bit rate  $R$ . The paper [20] provided a lower bound on  $h_{\text{est}}(\alpha, K)$  in terms of Lyapunov exponents under the assumption that the system preserves a smooth measure.

In [29], Matveev and Pogromsky studied three estimation objectives of increasing strength for discrete-time non-linear systems. For the weakest one, the smallest bit rate was again shown to be equal to the topological entropy. For the other ones, general upper and lower bounds were obtained which can be computed directly in terms of the linearized right-hand side of the equation generating the system. Moreover, a concrete design of a coding and decoding/estimation scheme was provided, which achieves the strongest state estimation objective with a bit rate described in terms of singular values of the linearization. Further related work for non-linear systems, mainly focusing on noiseless channels or erasure channels, includes [31, 46, 27, 53, 9, 42, 41, 23, 22].

There also have been important relevant studies in the information theory community. The source coding theory exclusively deals with the state estimation problem under information-rate constraints. Towards a further understanding of such a theory, we review Shannon's rate-distortion function, which is defined operationally as follows: Given a componentwise  $\mathbb{X}$ -valued stochastic process  $\{x_t\}_{t \in \mathbb{Z}_+}$ , a *rate-distortion* pair  $(R, D)$  is achievable if given the source process  $\{x_t\}$ , and a distortion function  $\rho : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}_+$ , there exist sequences of

1. Quantizers (Encoders):  $\mathcal{E}_n : \mathbb{X}^n \rightarrow \mathcal{M}(n)$  with  $|\mathcal{M}(n)| \leq 2^{Rn}$  and
2. Decoders:  $\mathcal{D}_n : \mathcal{M}(n) \rightarrow \mathbb{X}^n$  such that  $\hat{x}_t = \mathcal{D}_n(m)$ ,  $0 \leq t \leq n-1$ , with

$$D_n := \frac{1}{n} E \left[ \sum_{t=0}^{n-1} \rho(x_t, \hat{x}_t) \right] \leq D, \quad \lim_{n \rightarrow \infty} D_n \leq D.$$

The quantity  $\inf\{R : (R, D) \text{ is achievable}\}$  is the (operational) *rate-distortion* function of the source at the distortion level  $D$ . The case where  $D \rightarrow 0$  constitutes an important special setting. In the information theory community, the study of problems on state estimation has almost exclusively focused on stationary processes. For a class of such processes, the rate-distortion function admits a simple formula, often referred to as a *single-letter characterization*. We also note that if the source process is a finite-valued stationary and ergodic process, the rate-distortion function with  $D = 0$  reduces to the metric entropy of the source. As noted, there have been few contributions in the information theory literature on non-causal coding of non-stationary/unstable sources. These have only, to our knowledge, focused on Gaussian linear models: Consider the Gaussian auto-regressive (AR) process

$$x_t = - \sum_{k=1}^m a_k x_{t-k} + w_t,$$

where  $\{w_t\}$  is an i.i.d. zero-mean, Gaussian random sequence with variance  $E[w_t^2] = \sigma^2$ . If the roots of the complex polynomial  $H(z) = 1 + \sum_{k=1}^m a_k z^{-k}$  are all in the interior of the unit disk, then the process is asymptotically stationary and its rate-distortion function (with the distortion being the expected, normalized Euclidean error) is given parametrically by the following [14], obtained by considering the asymptotic distribution of the eigenvalues of the correlation matrix:

$$D_\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \min\left(\theta, \frac{1}{g(w)}\right) dw,$$

$$R(D_\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \max\left(\frac{1}{2} \log \frac{1}{\theta g(w)}, 0\right) dw,$$

where  $g(w) = \frac{1}{\sigma^2} |1 + \sum_{k=1}^m a_k e^{-ikw}|^2$ . If at least one root is on the unit circle or outside the unit disk, the analysis is more involved as the asymptotic eigenvalue distribution contains unbounded components. Gray and Hashimoto (see [14, 18, 16]) have shown, using the properties of the eigenvalues as well as Jensen's formula for integrations along the unit circle, that  $R(D_\theta)$  above should be replaced by

$$R(D_\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \max\left(\frac{1}{2} \log \frac{1}{\theta g(w)}, 0\right) dw + \sum_{k=1}^m \frac{1}{2} \max(0, \log(|\rho_k|^2)), \quad (2)$$

where  $\{\rho_k\}$  are the roots of the polynomial  $H$ . It is important to emphasize that the second term on the right-hand side of the equation is exactly the topological entropy for a linear system with eigenvalues  $\rho_k$ . On related problems, Berger [2] obtained the rate-distortion function for Wiener processes, and in addition, developed a two-part coding scheme, which was later generalized in [36] and [37] to unstable Markov processes driven by bounded noise.

It is useful to note that, when  $\mathbb{X}$  is finite, the source  $x_t$  is stationary, and  $D = 0$ , the solution to the operational problem stated above is given by the *entropy rate* of the source, i.e.  $\lim_{t \rightarrow \infty} H(x_t | x_{[0, t-1]})$ .

We also note that the above formulation is what is known as a non-causal construction. When one also inserts causality constraints, the problem typically becomes far more challenging, even when the source process is an i.i.d. process, because in this case the optimization problem becomes non-convex [50, Ch. 5]. We refer the reader to [26] and [50, Ch. 5] on further subtle differences between causal and non-causal coding of stochastic processes. Structural results, see e.g. [50, Ch. 10], [44, 43, 40, 48, 5] for finite horizon problems, and [25, 1, 45] for asymptotic average-distortion criteria, are available; in addition explicit bounds for the low-distortion regime building on Shannon's lower bound are available in the literature, see notably [26]. The findings in [26, 25, 1, 45] are particularly relevant to our study here, since they consider infinite-horizon criteria.

A related problem is the control of stochastic non-linear systems over communication channels. This problem has been studied in few publications, and mainly only for deterministic systems or deterministic channels. Recently, [49] studied stochastic stability properties such as asymptotic mean stationarity, ergodicity and non-positive normalized state entropy for a more general class of stochastic non-linear systems. The analysis in [49] builds on information-theoretic bounds and Markov-chain-theoretic constructions, however these bounds do not distinguish between the unstable and stable components of the tangent space associated with a dynamical non-linear system, except for the linear system case. Our paper here provides such a refinement, but only for estimation problems and in the low-distortion regime.

We also mention a different approach that can be found in [33]. In this paper, a general theory of non-stochastic information was developed, and for exponential state estimation with a given exponent  $\alpha > 0$  over an uncertain channel, a characterization of the critical bit rate similar to (1) was obtained.

## 1.1 Contributions

In the present paper, we consider stochastic non-linear systems of the form  $x_{t+1} = f(x_t, w_t)$  with  $(w_t)_{t \in \mathbb{Z}_+}$  being an i.i.d. sequence of random variables distributed according to a common probability  $\nu$ , modeling the noise. The initial state  $x_0$  is assumed to be a random variable, independent of the noise, distributed with  $x_0 \sim \pi_0$ . We assume that the system is estimated via a noisy channel with feedback of a very general form (a so-called Class A type channel, cf. [50]). At time  $t$ , the estimator generates an estimate  $\hat{x}_t$  of the state  $x_t$ , using the information it has received through the channel up to this time. We consider three estimation objectives of different strength for a given  $\epsilon > 0$ , describing the accuracy of the estimation:

- (E1) Eventual almost sure stability of the estimation error:  $d(x_t, \hat{x}_t) \leq \epsilon$  for all  $t \geq T(\epsilon)$  almost surely.
- (E2) Asymptotic almost sure stability of the estimation error:  $P(\limsup_{t \rightarrow \infty} d(x_t, \hat{x}_t) \leq \epsilon) = 1$ .
- (E3) Asymptotic quadratic stability in expectation:  $\limsup_{t \rightarrow \infty} E[d(x_t, \hat{x}_t)^2] \leq \epsilon$ .

We study the smallest channel capacity  $C_\epsilon$  above which a coder and a decoder/estimator can be designed achieving one of these objectives, and in particular  $C_0 = \lim_{\epsilon \downarrow 0} C_\epsilon$ .

In the case when the channel is noiseless, we view the state process as a shift dynamical system on the space of trajectories and we relate the capacity of the channel with the topological entropy of this shift restricted to the support of the stochastic process. This approach is new, to our knowledge, for the study of such stochastic processes.

In particular, we prove that  $C_0$  for the objectives (E1) and (E2) is bounded below by the topological or metric entropy of a shift dynamical system on the space of (typical) trajectories of the system, respectively. This entropy is typically infinite, because the noise in the system generates so many possible trajectories that the space of relevant trajectories becomes infinite-dimensional. The result, in the noise-free source and channel case, reduces to the existing results in the literature for deterministic setups.

The information-theoretic approach allows for results that do not require a compact state space or strong continuity conditions on the dynamical system. We show that essentially if the entropy rate of the source process is positive,  $C_0$  cannot be finite for asymptotically small estimation errors for all three estimation objectives. We also establish connections with high-rate quantization theory.

Another approach considered in the present paper views the given system as a random dynamical system whose base space consists of all noise realizations. Through this approach, we can show that for compact state spaces and noiseless channels,  $C_0$  is bounded below by the entropy of the random dynamical system for (E1) and (E2). Essentially, this implies that positive Lyapunov exponents are an obstruction to a state estimation with zero capacity. This is in correspondence with the results of the deterministic theory, but in view of the results described above, exhibits this approach as non-adequate for the problem, since the lower bounds obtained here are finite under mild assumptions.

Finally, we prove that for noiseless dynamical systems with finite topological entropy, state estimation over a discrete memoryless channel of finite capacity is possible for (E2), and in this case  $C_0$  is bounded above by the topological entropy. However, the coding and estimation policy used to prove this result is highly impractical to implement. We show that for linear systems and erasure channels, a more realistic coding and estimation policy exists.

The zero-noise results in this paper essentially coincide with the results in the literature on deterministic setups for either when the system noise is absent or the channel noise is absent. Our findings are also new in the information theory literature, where the estimation error results have exclusively focused on stable sources, except for linear models, as reviewed above.

One main message conveyed by our results is that typically none of above estimation objectives can be achieved over a channel of finite capacity with an arbitrarily small estimation error, i.e.,  $C_0 = \infty$ . More specifically, this is the case whenever the noise in the system influences the dynamics to such an extent that a recovery of the state with some accuracy  $\epsilon$  would allow a recovery of a sufficiently large part of the noise realization with an accuracy of the same order. This is made precise in several theorems, using principally different methodologies.

The present paper is organized as follows. Section 2 introduces some concepts and notations from dynamical systems and information theory and provides a detailed exposition of the state estimation problems studied in the subsequent sections. In Section 3 we derive lower bounds for the almost sure estimation objectives in terms of the entropy of an associated dynamical system on the space of trajectories, assuming a noiseless channel. The subsequent Section 4 is dedicated to the information-theoretic approach. Here we assume that the state space is  $\mathbb{R}^N$  and impose no restrictions on the channel. For all three estimation objectives, we formulate sufficient conditions for  $C_0 = \infty$ . Furthermore, for positive  $\epsilon$ , finite lower bounds for  $C_\epsilon$  are derived. Section 5 contains the lower bound result obtained via the random dynamical systems view. In Section 6, it is shown that a memoryless channel of finite capacity is sufficient to achieve (E2) if there is no noise in the system. Finally, in Section 7 some concluding remarks are presented.

## 2 Preliminaries

**Notation and definitions.** All logarithms in this paper are taken to the base 2.

If  $f : X \rightarrow Y$  is a measurable map between measurable spaces  $(X, \mathcal{F})$  and  $(Y, \mathcal{G})$ , we write  $f_*$  for the push-forward operator associated with  $f$  on the space of measures on  $(X, \mathcal{F})$ , i.e., for any measure  $\mu$  on  $(X, \mathcal{F})$ ,  $f_*\mu$  is the measure on  $(Y, \mathcal{G})$  defined by  $(f_*\mu)(G) := \mu(f^{-1}(G))$  for all  $G \in \mathcal{G}$ . Using this notation, the invariance of a measure  $\mu$  under a map  $f : X \rightarrow X$  can be expressed by writing  $f_*\mu = \mu$ .

An important concept used in this paper is the topological entropy of a dynamical system. If  $f : X \rightarrow X$  is a continuous map on a metric space  $(X, d)$ , and  $K \subset X$  is a compact set, we say that  $E \subset K$  is  $(n, \epsilon; f)$ -separated for some  $n \in \mathbb{N}$  and  $\epsilon > 0$  if for all  $x, y \in E$  with  $x \neq y$ ,  $d(f^i(x), f^i(y)) > \epsilon$  for some  $i \in \{0, 1, \dots, n-1\}$ . We write  $r_{\text{sep}}(n, \epsilon, K; f)$  for the maximal cardinality of an  $(n, \epsilon; f)$ -separated subset of  $K$  and define the topological entropy  $h_{\text{top}}(f, K)$  of  $f$  on  $K$  by

$$h_{\text{sep}}(f, \epsilon; K) := \limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{sep}}(n, \epsilon, K; f), \quad h_{\text{top}}(f, K) := \lim_{\epsilon \downarrow 0} h_{\text{sep}}(f, \epsilon; K).$$

If  $X$  is compact and  $K = X$ , we also omit the argument  $K$  and call  $h_{\text{top}}(f)$  the topological entropy of  $f$ . Alternatively, one can define  $h_{\text{top}}(f, K)$  using  $(n, \epsilon)$ -spanning sets. A set  $F \subset X$   $(n, \epsilon)$ -spans another set  $K \subset X$  if for each  $x \in K$  there is  $y \in F$  with  $d(f^i(x), f^i(y)) \leq \epsilon$  for  $i = 0, 1, \dots, n-1$ . Letting  $r_{\text{span}}(n, \epsilon, K; f)$  (or  $r_{\text{span}}(n, \epsilon, K)$  if the map  $f$  is clear from the context) denote the minimal cardinality of a set which  $(n, \epsilon)$ -spans  $K$ , the topological entropy of  $f$  on  $K$  satisfies

$$h_{\text{top}}(f, K) = \lim_{\epsilon \downarrow 0} \limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{span}}(n, \epsilon, K; f).$$

If  $f : X \rightarrow X$  is a measure-preserving map on a probability space  $(\Omega, \mathcal{F}, \mu)$ , i.e.,  $f_*\mu = \mu$ , its metric entropy  $h_\mu(f)$  is defined as follows. Let  $\mathcal{A}$  be a finite measurable partition of  $X$ . Then the entropy of  $f$  with respect to  $\mathcal{A}$  is defined by

$$h_\mu(f; \mathcal{A}) := \lim_{n \rightarrow \infty} \frac{1}{n} H_\mu \left( \bigvee_{i=0}^{n-1} f^{-i} \mathcal{A} \right). \quad (3)$$

Here  $\bigvee$  denotes the join operation, i.e.,  $\bigvee_{i=0}^{n-1} f^{-i} \mathcal{A}$  is the partition of  $X$  consisting of all intersections of the form  $A_0 \cap f^{-1}(A_1) \cap \dots \cap f^{-n+1}(A_{n-1})$  with  $A_i \in \mathcal{A}$ . For any partition  $\mathcal{B}$  of  $X$ ,  $H_\mu(\mathcal{B}) = -\sum_{B \in \mathcal{B}} \mu(B) \log \mu(B)$ . The existence of the limit in (3) follows from a subadditivity argument. The metric entropy of  $f$  is then defined by

$$h_\mu(f) := \sup_{\mathcal{A}} h_\mu(f; \mathcal{A}),$$

where the supremum is taken over all finite measurable partitions  $\mathcal{A}$  of  $X$ . If  $f$  is continuous,  $X$  is compact metric and  $\mu$  is ergodic, there is an alternative characterization of  $h_\mu(f)$  due to Katok [19]:

For any  $n \in \mathbb{N}$ ,  $\epsilon > 0$  and  $\delta \in (0, 1)$  put

$$r_{\text{span}}(n, \epsilon, \delta) := \min \{ r_{\text{span}}(n, \epsilon; A) : A \subset X \text{ Borel, } \mu(A) \geq 1 - \delta \}.$$

Then for every  $\delta \in (0, 1)$  it holds that

$$h_\mu(f) = \lim_{\epsilon \downarrow 0} \limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{span}}(n, \epsilon, \delta).$$

Topological and metric entropy are related to each other via the variational principle: For a continuous map  $f : X \rightarrow X$  on a compact metric space  $X$ ,

$$h_{\text{top}}(f) = \sup_{\mu} h_\mu(f),$$

where the supremum is taken over all  $f$ -invariant Borel probability measures.

We will also use several concepts from information theory. Let  $x$  be an  $\mathbb{X}$ -valued random variable, where  $\mathbb{X}$  is countable. The *entropy* of  $x$  is defined as  $H(x) := -\sum_{z \in \mathbb{X}} p(z) \log_2(p(z))$  where  $p$  is the probability mass function (pmf) of the random variable  $x$ . If  $x$  is an  $\mathbb{R}^n$ -valued random variable, and the probability measure induced by  $x$  is absolutely continuous with respect to the Lebesgue measure, the (*differential*) *entropy* of  $x$  is defined by  $h(x) := -\int_{\mathbb{X}} p(x) \log_2(p(x)) dx$ , where  $p(\cdot)$  is the probability density function (pdf) of  $x$ .

The *mutual information* between a discrete (continuous) random variable  $x$ , and another discrete (continuous) random variable  $y$ , defined on a common probability space, is defined as  $I(x; y) := H(x) - H(x|y)$ , where  $H(x)$  is the entropy of  $x$  (differential entropy if  $x$  is a continuous random variable), and  $H(x|y)$  is the conditional entropy of  $x$  given  $y$  ( $h(x|y)$  is the conditional differential entropy if  $x$  is a continuous random variable). For more general settings including when the random variables are continuous, discrete or a mixture of the two, mutual information is defined as  $I(x; y) := \sup_{Q_1, Q_2} I(Q_1(x); Q_2(y))$ , where  $Q_1$  and  $Q_2$  are quantizers with finitely many bins (see [15, Ch. 5]). An important relevant result is the following. Let  $x$  be a random variable and  $Q$  be a quantizer applied to  $x$ . Then,  $H(Q(x)) = I(x; Q(x)) = h(x) - h(x|Q(x))$ . For a concise overview of relevant information-theoretic concepts, we refer the reader to [50, Ch. 5]. For a more complete coverage, see [11, 6]. When the realization  $x$  of a random variable  $x_t$  needs to be explicitly mentioned, the event  $x_t = x$  will be emphasized. We use the conditional probability (expectation) notation  $P_x(\cdot)$  ( $E_x[\cdot]$ ) to denote  $P(\cdot|x_0 = x)$  ( $E[\cdot|x_0 = x]$ ).

Throughout the paper, we assume that all random variables are modeled on a common probability space  $(\Omega, \mathcal{F}, P)$ .

**Stochastic networked systems and estimation objectives.** In this paper, we consider non-linear noisy systems given by an equation of the form

$$x_{t+1} = f(x_t, w_t). \quad (4)$$

Here  $x_t$  is the state at time  $t$  and  $(w_t)_{t \in \mathbb{Z}_+}$  is an i.i.d. sequence of random variables with common distribution  $w_t \sim \nu$ , modeling the noise. In general, we assume that

$$f : X \times W \rightarrow X$$

is a Borel measurable map, where  $X$  and  $W$  are Polish spaces, so that for any  $w \in W$  the map  $f(\cdot, w)$  is a homeomorphism of  $X$ . We further assume that  $x_0$  is a random variable on  $X$  with an associated probability measure  $\pi_0$ , independent of  $(w_t)_{t \in \mathbb{Z}_+}$ . We use the notations

$$f_w : X \rightarrow X, \quad f_w(x) = f(x, w), \quad f^x : W \rightarrow X, \quad f^x(w) = f(x, w).$$

This system is connected over a noisy channel with a finite capacity to an estimator, as shown in Fig. 1. The estimator has access to the information it has received through the channel. A source coder maps the source symbols (i.e., state values), to corresponding channel inputs. The channel inputs are transmitted through the channel; we assume that the channel is a discrete channel with input alphabet  $\mathcal{M}$  and output alphabet  $\mathcal{M}'$ .

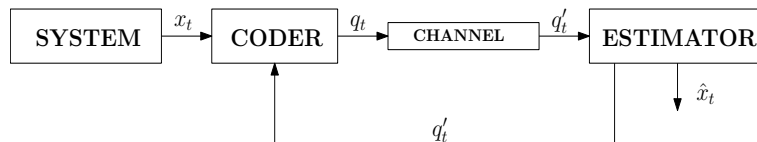


Figure 1: Estimation over a noisy channel with feedback

We refer by a *Coding Policy*  $\Pi$ , to a sequence of functions  $(\gamma_t^e)_{t \in \mathbb{Z}_+}$  which are causal such that the channel input at time  $t$ ,  $q_t \in \mathcal{M}$ , under  $\Pi$  is generated by a function of its local information, i.e.,

$$q_t = \gamma_t^e(\mathcal{I}_t^e),$$

where  $\mathcal{I}_t^e = \{x_{[0,t]}, q'_{[0,t-1]}\}$  and  $q_t \in \mathcal{M}$ , the channel input alphabet given by  $\mathcal{M} = \{1, 2, \dots, M\}$ , for  $0 \leq t \leq T-1$ . Here, we use the notation  $x_{[0,t-1]} = \{x_s, 0 \leq s \leq t-1\}$  for  $t \geq 1$ .

The channel maps  $q_t$  to  $q'_t$  in a stochastic fashion so that  $P(q'_t|q_t, q_{[0,t-1]}, q'_{[0,t-1]})$  is a conditional probability measure on  $\mathcal{M}'$  for all  $t \in \mathbb{Z}_+$ . If this expression is equal to  $P(q'_t|q_t)$ , the channel is said to be memoryless, i.e., the past variables do not affect the channel output  $q'_t$  given the current channel input  $q_t$ .

Even though many of our results will focus on discrete memoryless channels (DMC), we will have occasions to study more general channels. In particular, we consider the following class of channels (see [50, Def. 8.5.1]).

**Definition 2.1** *A channel is said to be of Class A type, if*

- *it satisfies the following Markov chain condition:*

$$q'_t \leftrightarrow q_t, q_{[0,t-1]}, q'_{[0,t-1]} \leftrightarrow \{x_0, w_s, s \geq 0\},$$

*i.e., almost surely, for all Borel sets  $B$ ,*

$$P(q'_t \in B | q_t, q_{[0,t-1]}, q'_{[0,t-1]}, x_0, w_s, s \geq 0) = P(q'_t \in B | q_t, q_{[0,t-1]}, q'_{[0,t-1]})$$

*for all  $t \geq 0$ , and*

- *its capacity with feedback is given by*

$$C = \lim_{T \rightarrow \infty} \max_{\{P(q_t|q_{[0,t-1]}, q'_{[0,t-1]}), 0 \leq t \leq T-1\}} \frac{1}{T} I(q_{[0,T-1]} \rightarrow q'_{[0,T-1]}),$$

*where the directed mutual information is defined by*

$$I(q_{[0,T-1]} \rightarrow q'_{[0,T-1]}) := \sum_{t=1}^{T-1} I(q_{[0,t]}; q'_t | q'_{[0,t-1]}) + I(q_0; q'_0).$$

Discrete noiseless channels and memoryless channels belong to this class; for such channels, feedback does not increase the capacity [6]. Class A type channels also include finite state stationary Markov channels which are indecomposable [34], and non-Markov channels which satisfy certain symmetry properties [7]. Further examples can be found in [39, 8].

The receiver, upon receiving the information from the channel, generates an estimate  $\hat{x}_t$  at time  $t$ , also causally: An admissible causal estimation policy is a sequence of functions  $(\gamma_t^d)_{t \in \mathbb{Z}_+}$  such that  $\hat{x}_t = \gamma_t^d(q'_{[0,t]})$  with

$$\gamma_t^d : (\mathcal{M}')^{t+1} \rightarrow X, \quad t \geq 0.$$

For a given  $\epsilon > 0$ , we denote by  $C_\epsilon$  the smallest channel capacity above which there exist an encoder and an estimator so that one of the following estimation objectives is achieved:

(E1) Eventual almost sure stability of the estimation error: There exists  $T(\epsilon) \geq 0$  so that

$$\sup_{t \geq T(\epsilon)} d(x_t, \hat{x}_t) \leq \epsilon \quad \text{a.s.} \tag{5}$$

(E2) Asymptotic almost sure stability of the estimation error:

$$P\left(\limsup_{t \rightarrow \infty} d(x_t, \hat{x}_t) \leq \epsilon\right) = 1. \quad (6)$$

(E3) Asymptotic quadratic stability of the estimation error in expectation:

$$\limsup_{t \rightarrow \infty} E[d(x_t, \hat{x}_t)^2] \leq \epsilon. \quad (7)$$

It is easy to see that (5) implies (6) and (7) (with  $\epsilon^2$ ). On the other hand, (6) and (7) do not imply one another in general; one can construct examples where one holds and the other does not.

The primary goal of the paper is to find  $C_\epsilon$  and in particular

$$C_0 := \lim_{\epsilon \downarrow 0} C_\epsilon. \quad (8)$$

Observe that this limit exists as a number in  $[0, \infty]$ , since  $C_\epsilon$  is non-decreasing as  $\epsilon \downarrow 0$ .

### 3 Bounds through a dynamical systems approach

In this section, we assume that the channel is noiseless. Under this assumption, we derive lower bounds of  $C_0$  for the estimation objectives (5) and (6).

We consider the space  $X^{\mathbb{Z}_+}$  of all sequences in  $X$ , equipped with the product topology. We write  $\bar{x} = (x_0, x_1, x_2, \dots)$  for the elements of  $X^{\mathbb{Z}_+}$  and we fix the product metric

$$D(\bar{x}, \bar{y}) := \sum_{t=0}^{\infty} \frac{1}{2^t} \frac{d(x_t, y_t)}{1 + d(x_t, y_t)},$$

where  $d(\cdot, \cdot)$  is the given metric on  $X$ . A natural dynamical system on  $X^{\mathbb{Z}_+}$  is the shift map  $\theta : X^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+}$ ,  $(\theta\bar{x})_t \equiv x_{t+1}$ , which is continuous with respect to the product topology. An analogous shift map is defined on  $W^{\mathbb{Z}_+}$  and denoted by  $\vartheta$ .

Observing that the sequence of random variables  $(x_t)_{t \in \mathbb{Z}_+}$  forms a Markov chain, when  $x_0$  is fixed, the following lemma shows how a stationary measure of this Markov chain defines an invariant measure for  $\theta$ .

**Lemma 3.1** *Let  $\pi$  be a stationary measure of the Markov chain  $(x_t)_{t \in \mathbb{Z}_+}$ . Then an invariant Borel probability measure  $\mu$  for  $\theta$  is defined on cylinder sets by*

$$\mu(B_0 \times B_1 \times \dots \times B_n \times X^{[n+1, \infty)}) := \int_{B_0 \times B_1 \times \dots \times B_n} \pi(dx_0) P(dx_1|x_0) P(dx_2|x_1) \dots P(dx_n|x_{n-1}),$$

where  $B_0, B_1, \dots, B_n$  are arbitrary Borel sets in  $X$ . Here

$$P(x_{n+1} \in B | x_n = x) = P(f(x_n, w) \in B | x_n = x) = \nu(\{w \in W : f(x, w) \in B\}).$$

The support of  $\mu$  is contained in the closure of the set of all trajectories, i.e.,

$$\text{supp } \mu \subset \text{cl } \mathcal{T}, \quad \mathcal{T} := \{\bar{x} \in X^{\mathbb{Z}_+} : \exists w_t \in W \text{ with } x_{t+1} \equiv f(x_t, w_t), t \in \mathbb{Z}_+\}.$$

**Proof:** We consider the map

$$G : X \times W^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+},$$

which maps a pair  $(x_0, \bar{w})$  with  $\bar{w} = (w_t)_{t \in \mathbb{Z}_+}$  to the trajectory  $(x_t)_{t \in \mathbb{Z}_+}$  obtained by  $x_{t+1} := f(x_t, w_t)$ . We claim that this map is measurable and its associated push-forward operator on measures maps  $\pi_0 \times \nu^{\mathbb{Z}_+}$  to  $\mu$ . To prove that  $G$  is measurable, consider a cylinder set  $A = B_0 \times \dots \times B_n \times X^{[n+1, \infty)}$  in  $X^{\mathbb{Z}_+}$ . Then

$$G^{-1}(A) = \{(x_0, \bar{w}) : x_0 \in B_0, G(x_0, \bar{w})_1 \in B_1, \dots, G(x_0, \bar{w})_n \in B_n\}.$$



Hence,  $G^{-1}(A)$  can be expressed as the preimage of  $B_0 \times \cdots \times B_n \subset X^{n+1}$  under the map

$$(x_0, \bar{w}) \mapsto (x_0, f_{w_0}(x_0), f_{w_1} \circ f_{w_0}(x_0), \dots, f_{w_{n-1}} \circ \cdots \circ f_{w_0}(x_0)).$$

To show that this map is measurable, it suffices to show that each component is a measurable map. This follows from the fact that the projection  $W^{\mathbb{Z}_+} \rightarrow W^{n+1}$  to the first  $n+1$  components is measurable and  $f$  is measurable. Hence, we have proved that  $G$  is measurable. To see that  $G_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \mu$ , observe that for a set of the form  $A = B_0 \times B_1 \times X^{[2, \infty)}$  we have

$$\begin{aligned} & \pi_0 \times \nu^{\mathbb{Z}_+}(\{(x_0, \bar{w}) : x_0 \in B_0, f_{w_0}(x_0) \in B_1\}) \\ &= \int_X \int_{W^{\mathbb{Z}_+}} \nu^{\mathbb{Z}_+}(d\bar{w}) \pi_0(dx_0) \mathbf{1}_{B_0}(x_0) \mathbf{1}_{B_1}(f_{w_0}(x_0)) \\ &= \int_{B_0} \pi_0(dx_0) \int_W \nu(dw) \mathbf{1}_{B_1}(f_w(x_0)) = \int_{B_0} \pi_0(dx_0) \nu(\{w \in W : f_w(x_0) \in B_1\}) \\ &= \mu(B_0 \times B_1 \times X^{[2, \infty)}). \end{aligned} \tag{9}$$

For more general cylinder sets, the claim follows inductively. The fact that  $\text{supp } \mu$  is contained in  $\text{cl } \mathcal{T}$  follows from

$$\begin{aligned} \mu(\text{cl } \mathcal{T}) &= G_*[\pi_0 \times \nu^{\mathbb{Z}_+}](\text{cl } \mathcal{T}) = \pi_0 \times \nu^{\mathbb{Z}_+}(G^{-1}(\text{cl } \mathcal{T})) \geq \pi_0 \times \nu^{\mathbb{Z}_+}(G^{-1}(\mathcal{T})) \\ &= \pi_0 \times \nu^{\mathbb{Z}_+}(G^{-1}(G(X \times W^{\mathbb{Z}_+}))) = \pi_0 \times \nu^{\mathbb{Z}_+}(X \times W^{\mathbb{Z}_+}) = 1. \end{aligned}$$

Finally, we show that  $\mu$  is  $\theta$ -invariant. To this end, note that the map  $\Phi : X \times W^{\mathbb{Z}_+} \rightarrow X \times W^{\mathbb{Z}_+}$ ,  $(x, \bar{w}) \mapsto (f(x, w_0), \vartheta \bar{w})$ , satisfies  $\theta \circ G = G \circ \Phi$ . Using that

$$\begin{aligned} \pi_0 \times \nu^{\mathbb{Z}_+}(\Phi^{-1}(A \times B)) &= \pi_0 \times \nu^{\mathbb{Z}_+}(\{(x_0, \bar{w}) : f_{w_0}(x_0) \in A, \vartheta \bar{w} \in B\}) \\ &= \pi_0 \times \nu^{\mathbb{Z}_+}\left(\bigcup_{x_0 \in X} \{x_0\} \times ((f^{x_0})^{-1}(A) \times B)\right) \\ &= \int_X \pi_0(dx_0) \nu(\{w : f(x_0, w) \in A\}) \nu^{\mathbb{Z}_+}(B) \\ &= \nu^{\mathbb{Z}_+}(B) \int_X \pi_0(dx) P(x, A) = \pi_0(A) \nu^{\mathbb{Z}_+}(B) = \pi_0 \times \nu^{\mathbb{Z}_+}(A \times B), \end{aligned}$$

i.e.,  $\Phi_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \pi_0 \times \nu^{\mathbb{Z}_+}$ , we find that

$$\theta_* \mu = \theta_* G_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = G_* \Phi_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = G_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \mu,$$

completing the proof.  $\square$

The next lemma will be useful to relate  $C_0$  for the estimation objective (5) to the topological entropy of the shift map  $\theta$ .

**Lemma 3.2** *Let  $f : X \rightarrow X$  be a homeomorphism on a compact metric space  $(X, d)$ . Fix  $\epsilon > 0$  and  $n_0 \in \mathbb{N}$ . For  $n > n_0$  we say that a set  $E \subset X$  is  $(n, \epsilon; n_0)$ -separated if  $d(f^i(x), f^i(y)) > \epsilon$  for some  $i \in \{n_0, n_0 + 1, \dots, n - 1\}$ , whenever  $x, y \in E$  with  $x \neq y$ . We write  $r_{\text{sep}}(n, \epsilon; n_0, f)$  for the maximal cardinality of an  $(n, \epsilon; n_0)$ -separated set. Then, for any choice of  $n_0(\epsilon) \in \mathbb{N}$ ,  $\epsilon > 0$ , we have*

$$h_{\text{top}}(f) = \lim_{\epsilon \downarrow 0} \limsup_{n_0(\epsilon) < n \rightarrow \infty} \frac{1}{n} \log r_{\text{sep}}(n, \epsilon; n_0(\epsilon)). \tag{10}$$

**Proof:** Any  $(n, \epsilon; n_0(\epsilon))$ -separated set is trivially  $(n, \epsilon)$ -separated, hence  $r_{\text{sep}}(n, \epsilon) \geq r_{\text{sep}}(n, \epsilon; n_0(\epsilon))$ , implying the inequality “ $\geq$ ” in (10). Conversely, assume that  $E$  is  $(n, \epsilon)$ -separated and put  $E' := f^{-n_0(\epsilon)}(E)$ . Then  $|E'| = |E|$  and  $E'$  is  $(n_0(\epsilon) + n, \epsilon; n_0(\epsilon))$ -separated. This implies

$$\frac{n + n_0(\epsilon)}{n} \frac{1}{n + n_0(\epsilon)} \log r_{\text{sep}}(n_0(\epsilon) + n, \epsilon; n_0(\epsilon)) \geq \frac{1}{n} \log r_{\text{sep}}(n, \epsilon).$$

Letting  $n \rightarrow \infty$  on both sides, we find that

$$\limsup_{n_0(\epsilon) < n \rightarrow \infty} \frac{1}{n} \log r_{\text{sep}}(n, \epsilon; n_0(\epsilon)) \geq h_{\text{sep}}(f, \epsilon).$$

Finally, letting  $\epsilon \downarrow 0$ , the desired inequality follows.  $\square$

**Theorem 3.1** *Consider the estimation objective (5) for an initial measure  $\pi_0$  which is stationary under the Markov chain  $(x_t)_{t \in \mathbb{Z}_+}$ . If  $\text{supp } \mu$  is not compact, we have  $C_0 = \infty$ . Otherwise,*

$$C_0 \geq h_{\text{top}}(\theta|_{\text{supp } \mu}).$$

*As a consequence, the metric entropy of  $\theta$  with respect to  $\mu$  is also a lower bound of  $C_0$ .*

**Proof:** Assume that for some  $\epsilon > 0$  the objective (5) is achieved by a pair of coder and estimator via a noiseless channel of capacity  $C = \log |\mathcal{M}|$ . Then for every  $k \in \mathbb{N}$  we define the set

$$\mathcal{E}_k := \{(\hat{x}_0, \hat{x}_1, \dots, \hat{x}_{k-1}) : q_t \in \mathcal{M}, 0 \leq t \leq k-1\}$$

of all possible estimation sequences of length  $k$  the estimator can generate in the time interval  $[0, k-1]$ .

Assume to the contrary that there exists a measurable set  $A \subset X^{\mathbb{Z}_+}$  of positive measure  $\alpha := \mu(A) > 0$  so that for every  $\bar{x} = (x_t)_{t \in \mathbb{Z}_+} \in A$  there exists  $t \geq T(\epsilon)$  with  $d(x_t, \hat{x}_t) > \epsilon$  in case the sequence  $(x_t)$  is realized as a trajectory of the system. If  $G : X \times W^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+}$  is the map from the proof of Lemma 3.1, then the preimage  $G^{-1}(A)$  is measurable in  $X \times W^{\mathbb{Z}_+}$  with  $\pi_0 \times \nu^{\mathbb{Z}_+}$ -measure  $\alpha > 0$ . This contradicts the assumption that the almost sure estimation objective (5) is achieved. Hence, the set

$$\tilde{\mathcal{T}} := \{\bar{x} \in X^{\mathbb{Z}_+} : d(x_t, \hat{x}_t) \leq \epsilon \text{ for all } t \geq T(\epsilon)\}$$

has measure one and consequently is dense in  $\text{supp } \mu$ .

Choose  $\tau = \tau(\epsilon)$  large enough so that

$$\sum_{t=\tau}^{\infty} \frac{1}{2^t} \leq \epsilon.$$

Let  $E \subset \text{supp } \mu$  be a finite  $(k, 5\epsilon; T(\epsilon))$ -separated set for some  $k > T(\epsilon)$ . Since  $\tilde{\mathcal{T}}$  is dense in  $\text{supp } \mu$ , a small perturbation of  $E$  yields a  $(k, 5\epsilon; T(\epsilon))$ -separated set in  $\tilde{\mathcal{T}}$  with the same cardinality as  $E$  (using that  $\theta$  is continuous). Hence, we may assume  $E \subset \tilde{\mathcal{T}}$ . We define a map  $\alpha : E \rightarrow \mathcal{E}_{k+\tau}$  by assigning to  $(x_t)_{t \in \mathbb{Z}_+} \in E$  the estimation sequence generated by the estimator when it receives the signals  $q_t = q_t(x_0, \dots, x_t)$  for  $t = 0, 1, \dots, k + \tau - 1$ .

Assuming  $\alpha(\bar{x}) = \alpha(\bar{y})$  for some  $\bar{x}, \bar{y} \in E$ , we find for  $T(\epsilon) \leq t \leq k$  that

$$\begin{aligned} D(\theta^t(\bar{x}), \theta^t(\bar{y})) &\leq \sum_{s=0}^{\tau-1} \frac{1}{2^s} \frac{d(x_{t+s}, y_{t+s})}{1 + d(x_{t+s}, y_{t+s})} + \sum_{s=\tau}^{\infty} \frac{1}{2^s} \\ &\leq \sum_{s=0}^{\tau-1} \frac{1}{2^s} d(x_{t+s}, \hat{x}_{t+s}) + \sum_{s=0}^{\tau-1} \frac{1}{2^s} d(\hat{y}_{t+s}, y_{t+s}) + \epsilon \leq 2\epsilon + 2\epsilon + \epsilon = 5\epsilon, \end{aligned}$$

implying  $\bar{x} = \bar{y}$ , since  $E$  is  $(k, 5\epsilon; T(\epsilon))$ -separated. Hence, the map  $\alpha$  is injective.

The set  $\text{supp } \mu$  is a closed subset of the complete metric space  $(X^{\mathbb{Z}_+}, D)$ , hence it is also a complete metric space. If we assume that  $\text{supp } \mu$  is not compact, it thus follows that  $\text{supp } \mu$  is not totally bounded, implying that  $(k, 5\epsilon; T(\epsilon))$ -separated subsets of  $\text{supp } \mu$  of arbitrarily large (finite) cardinality exist. Hence,  $\mathcal{E}_{k+\tau}$  must be infinite, leading to the contradiction  $|\mathcal{M}| = \infty$ . Hence, in this case the estimation problem cannot be solved via a channel of finite capacity.

Now assume that  $\text{supp } \mu$  is compact. Choosing a maximal  $(k, 5\epsilon; T(\epsilon))$ -separated set  $E$ , for the dynamical system  $\theta|_{\text{supp } \mu} : \text{supp } \mu \rightarrow \text{supp } \mu$  we obtain the inequality

$$r_{\text{sep}}(k, 5\epsilon; T(\epsilon)) \leq |\mathcal{E}_{k+\tau}| \leq |\mathcal{M}|^{k+\tau}.$$

This implies

$$\limsup_{k \rightarrow \infty} \frac{1}{k} \log r_{\text{sep}}(k, 5\epsilon; T(\epsilon)) \leq \log |\mathcal{M}| = C.$$

Using Lemma 3.2, the result follows by letting  $C \rightarrow C_\epsilon$  and  $\epsilon \downarrow 0$ . The statement about the metric entropy now follows from the variational principle.  $\square$

**Remark 3.1** *To make the statement of the theorem clearer, let us consider the two extreme cases when there is no noise and when there is only noise:*

- (i) *If the system is deterministic, i.e.,  $x_{t+1} = f(x_t)$  for a homeomorphism  $f : X \rightarrow X$  of a compact metric space  $X$ , then  $\pi_0$  is an invariant measure of  $f$ . Moreover,  $P(x_t \in B | x_{t-1} = x) = 1$  if  $f(x) \in B$  and 0 otherwise, implying*

$$\mu(B_0 \times B_1 \times \cdots \times B_n \times X^{[n+1, \infty)}) = \pi_0(B_0 \cap f^{-1}(B_1) \cap f^{-2}(B_2) \cap \cdots \cap f^{-n}(B_n)).$$

*From this expression, we see that the support of  $\mu$  is contained in the set  $\mathcal{T}$  of all trajectories of  $f$  (which in this case coincides with its closure), as already proved in Lemma 3.1. The map  $h : \mathcal{T} \rightarrow X$  defined by  $h(\bar{x}) := x_0$ , is easily seen to be a homeomorphism, which conjugates  $\theta|_{\mathcal{T}}$  and  $f$ . That is, the following diagram commutes:*

$$\begin{array}{ccc} \mathcal{T} & \xrightarrow{\theta} & \mathcal{T} \\ h \downarrow & & \downarrow h \\ X & \xrightarrow{f} & X \end{array}$$

*Since  $h_*\mu = \pi_0$  and conjugate systems have the same entropy, our theorem implies*

$$C_0 \geq h_{\text{top}}(f; \text{supp } \pi_0).$$

*The right-hand side of this inequality is finite under mild assumptions, e.g., if  $f$  is Lipschitz continuous on  $\text{supp } \pi_0$  and  $\text{supp } \pi_0$  has finite upper box dimension (see [4]). These conditions are in particular satisfied when  $f$  is a diffeomorphism on a finite-dimensional manifold. However, one should be aware that even on a compact interval there exist continuous maps with infinite topological entropy on the support of an invariant measure.*

- (ii) *Assume that  $X = W$  is compact and the system is given by  $x_{t+1} = w_t$ , i.e., the trajectories are only determined by the noise. In this case, with  $\pi_0 := \nu$ , the measure  $\mu$  is the product measure  $\nu^{\mathbb{Z}_+}$ . Hence,  $C_0$  is bounded below by the topological entropy of the shift on  $W^{\mathbb{Z}_+}$  restricted to  $\text{supp } \nu^{\mathbb{Z}_+} = (\text{supp } \nu)^{\mathbb{Z}_+}$ . This number is finite if and only if  $\text{supp } \nu$  is finite and in this case is given by  $\log |\text{supp } \nu|$ .*

If the system is not deterministic, then usually  $C_0 = \infty$ . In fact, this is always the case if the estimator is able to recover the noise to a sufficiently large extent. The following corollary treats the case, when the noise can be recovered completely from the state trajectory.

**Corollary 3.1** *Additionally to the assumptions in Theorem 3.1, suppose that  $W$  and  $X$  are compact and  $f^x : W \rightarrow X$  is invertible for every  $x \in X$  so that  $(x, y) \mapsto (f^x)^{-1}(y)$  is continuous. Then*

$$C_0 \geq h_{\text{top}}(\Phi|_{\text{supp}(\pi_0 \times \nu^{\mathbb{Z}_+})}) \geq h_{\text{top}}(\vartheta|_{\text{supp } \nu^{\mathbb{Z}_+}}), \quad (11)$$

*where  $\Phi : X \times W^{\mathbb{Z}_+} \rightarrow X \times W^{\mathbb{Z}_+}$  is the skew-product map  $(x, \bar{w}) \mapsto (f_{w_0}(x), \vartheta \bar{w})$ . As a consequence,  $C_0 = \infty$  whenever  $\text{supp } \nu$  contains infinitely many elements.*

**Proof:** We consider the map  $h : X^{\mathbb{Z}_+} \rightarrow W^{\mathbb{Z}_+}$ ,  $\bar{x} \mapsto \bar{w} = (w_t)_{t \in \mathbb{Z}_+}$  with

$$w_t = (f^{x_t})^{-1}(x_{t+1}).$$

If we equip  $W^{\mathbb{Z}_+}$  with the product topology,  $h$  becomes continuous. Indeed, if the distance of two points  $\bar{x}^1, \bar{x}^2 \in X^{\mathbb{Z}_+}$  is small, then the distances  $d_X(\bar{x}_t^1, \bar{x}_t^2)$  are small for finitely many values of  $t$ . Hence, by the uniform continuity of  $(x, y) \mapsto f_x^{-1}(y)$  on the compact space  $X \times X$ , also the distances  $d_W(h(\bar{x}^1)_t, h(\bar{x}^2)_t)$  can be made small for sufficiently many values of  $t$ , guaranteeing that  $D(h(\bar{x}^1), h(\bar{x}^2))$  becomes small, where  $D$  is a product metric on  $W^{\mathbb{Z}_+}$ .

The map  $G : X \times W^{\mathbb{Z}_+} \rightarrow X^{\mathbb{Z}_+}$ , used in the proof of Lemma 3.1, satisfies

$$h(G(x_0, \bar{w})) = \bar{w} \quad \text{for all } (x_0, \bar{x}) \in X \times W^{\mathbb{Z}_+},$$

because we can write

$$G(x_0, \bar{w}) = (x_0, f^{x_0}(w_0), f^{x_1}(w_1), f^{x_2}(w_2), \dots).$$

Consequently,  $G$  - as a map from  $X \times W^{\mathbb{Z}_+}$  to the space  $\mathcal{T}$  of trajectories - is invertible with

$$G^{-1}(\bar{x}) = (x_0, h(\bar{x})).$$

From the assumptions it follows that  $G$  is continuous, hence  $G$  is a homeomorphism and  $\mathcal{T}$  is compact. By the proof of Lemma 3.1, we have  $\theta \circ G = G \circ \Phi$ , where  $\Phi$  is the skew-product map  $\Phi(x, \bar{w}) = (f(x, w_0), \vartheta \bar{w})$  and  $G_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \mu$ . Hence,  $G$  is a topological conjugacy between  $\theta|_{\text{supp } \mu}$  and  $\Phi|_{\text{supp}(\pi_0 \times \nu^{\mathbb{Z}_+})}$ , implying

$$C_0 \geq h_{\text{top}}(\theta|_{\text{supp } \mu}) = h_{\text{top}}(\Phi|_{\text{supp}(\pi_0 \times \nu^{\mathbb{Z}_+)}).$$

Since the projection map  $\pi : (x, \bar{w}) \mapsto \bar{w}$  exhibits  $\vartheta$  as a topological factor of  $\Phi$  and  $\pi_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \nu^{\mathbb{Z}_+}$ , the second inequality in (11) follows.  $\square$

**Example 3.1** Let  $X = W = S^1 = \mathbb{R}/\mathbb{Z}$ . Let  $f(x, w) = x + w \bmod 1$  and let  $\pi_0 = \nu$  be the normalized Lebesgue measure on  $S^1$ . In this case, the map  $f^x : S^1 \rightarrow S^1$ ,  $w \mapsto x + w$ , is obviously invertible and  $(x, y) \mapsto (f^x)^{-1}(y) = x - y$  is continuous. Hence,  $C_0 = \infty$  for the estimation objective (5).

Now we consider the asymptotic estimation objective (6).

**Theorem 3.2** Consider the estimation objective (6) for an initial measure  $\pi_0$  which is stationary and ergodic under the Markov chain  $(x_t)_{t \in \mathbb{Z}_+}$ . Then, if  $\text{supp } \mu$  is compact,

$$C_0 \geq h_\mu(\theta).$$

**Proof:** First observe that the ergodicity of  $\pi_0$  implies the ergodicity of  $\mu$ . Indeed, it is well-known that the product measure  $\pi_0 \times \nu^{\mathbb{Z}_+}$  is ergodic for the skew-product  $\Phi$  if  $\pi_0$  is ergodic (cf. [21]). Since  $G_*(\pi_0 \times \nu^{\mathbb{Z}_+}) = \mu$  and  $\theta \circ G = G \circ \Phi$ , this implies the ergodicity of  $\mu$ . Now consider a noiseless channel with input alphabet  $\mathcal{M}$  and a pair of coder and decoder/estimator which solves the estimation problem (6) for some  $\epsilon > 0$ . For every  $\bar{x} \in X^{\mathbb{Z}_+}$  and  $\delta > \epsilon$  let

$$T(\bar{x}, \delta) := \inf \left\{ k \in \mathbb{N} : \sup_{t \geq k} d(x_t, \hat{x}_t) \leq \delta \right\},$$

where the infimum is defined as  $+\infty$  if the corresponding set is empty. Note that  $T(\bar{x}, \delta)$  depends measurably on  $\bar{x}$ . Define

$$B^K(\delta) := \{ \bar{x} \in \text{supp } \mu : T(\bar{x}, \delta) \leq K \} \quad \text{for all } \delta > \epsilon \text{ and } K \in \mathbb{N},$$

and observe that these sets are measurable. From (6) it follows that for every  $\delta > \epsilon$ ,

$$\lim_{K \rightarrow \infty} \mu(B^K(\delta)) = \mu\left(\bigcup_{K \in \mathbb{N}} B^K(\delta)\right) = 1.$$

Fixing a  $K$  large enough so that  $\mu(B^K(\delta)) > 0$ , Katok's characterization of metric entropy yields the assertion, which is proved with the same arguments as in the proof of Theorem 3.1, using the simple fact a maximal  $(n, \epsilon)$ -separated set contained in some set  $K$  also  $(n, \epsilon)$ -spans  $K$ . For more details, see the proof of Theorem 5.2.  $\square$

## 4 An information-theoretic view

In this section, we again allow noise in the channel and assume that  $X = \mathbb{R}^N$ . We prove further impossibility results via information-theoretic methods. These results will shed more light on the precise conditions that force  $C_0$  to be infinite.

**Theorem 4.1** *Consider system (4) with state space  $X = \mathbb{R}^N$ . Suppose that*

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{t-1}) > 0$$

and  $h(x_t) < \infty$  for all  $t \in \mathbb{Z}_+$ . Then, under (7) (and thus under (5)),

$$C_\epsilon \geq \left( \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{t-1}) \right) - \frac{N}{2} \log(2\pi e \epsilon).$$

In particular,  $C_0 = \infty$ .

**Proof:** Let  $(\epsilon_t)_{t \in \mathbb{Z}_+}$  be a sequence of non-negative real numbers so that  $E[\|x_t - \hat{x}_t\|^2] \leq \epsilon_t$  for all  $t \in \mathbb{Z}_+$  and  $\limsup_{t \rightarrow \infty} \epsilon_t \leq \epsilon$ . Observe that for  $t > 0$ :

$$\begin{aligned} I(q'_t; q_{[0,t]} | q'_{[0,t-1]}) &= H(q'_t | q'_{[0,t-1]}) - H(q'_t | q_{[0,t]}, q'_{[0,t-1]}) \\ &= H(q'_t | q'_{[0,t-1]}) - H(q'_t | q_{[0,t]}, x_t, q'_{[0,t-1]}) \\ &\geq H(q'_t | q'_{[0,t-1]}) - H(q'_t | x_t, q'_{[0,t-1]}) \\ &= I(x_t; q'_t | q'_{[0,t-1]}). \end{aligned} \tag{12}$$

Here, (12) follows from the assumption that the channel is of Class A type. Define

$$R_T := \max_{\{P(q_t | q_{[0,t-1]}, q'_{[0,t-1]}), 0 \leq t \leq T-1\}} \frac{1}{T} \sum_{t=0}^{T-1} I(q'_t; q_{[0,t]} | q'_{[0,t-1]}).$$

Now consider the following:

$$\begin{aligned} \lim_{T \rightarrow \infty} R_T &\geq \limsup_{T \rightarrow \infty} \frac{1}{T} \left( \sum_{t=1}^{T-1} I(x_t; q'_t | q'_{[0,t-1]}) + I(x_0; q'_0) \right) \\ &= \limsup_{T \rightarrow \infty} \frac{1}{T} \left( \sum_{t=1}^{T-1} \left( h(x_t | q'_{[0,t-1]}) - h(x_t | q'_{[0,t]}) \right) + I(x_0; q'_0) \right) \\ &\geq \limsup_{T \rightarrow \infty} \frac{1}{T} \left( \sum_{t=1}^{T-1} \left( h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - h(x_t | q'_{[0,t]}) \right) + I(x_0; q'_0) \right) \\ &= \limsup_{T \rightarrow \infty} \frac{1}{T} \left( \sum_{t=1}^{T-1} \left( h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - h(x_t - \hat{x}_t | q'_{[0,t]}) \right) + I(x_0; q'_0) \right) \\ &\geq \limsup_{T \rightarrow \infty} \frac{1}{T} \left( \sum_{t=1}^{T-1} \left( h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - h(x_t - \hat{x}_t) \right) + I(x_0; q'_0) \right) \\ &\geq \limsup_{T \rightarrow \infty} \frac{1}{T} \left( \sum_{t=1}^{T-1} \left( h(x_t | x_{[0,t-1]}, q'_{[0,t-1]}) - \frac{N}{2} \log(2\pi e \epsilon_t) \right) + I(x_0; q'_0) \right) \end{aligned} \tag{13}$$

$$= \limsup_{T \rightarrow \infty} \frac{1}{T} \left( \sum_{t=1}^{T-1} h(x_t | x_{t-1}) \right) - \frac{N}{2} \log(2\pi e \epsilon). \quad (14)$$

Here, the second inequality uses the property that entropy decreases under conditioning on more information. The second equality follows from the fact that  $\hat{x}_t$  is a function of  $q'_{[0,t]}$ , and the last inequality follows from that fact that among all real random variables  $X$  that satisfy a given second moment constraint  $E[X^2] \leq \epsilon$ , a Gaussian maximizes the entropy and the differential entropy in this case is given by  $\frac{1}{2} \log(2\pi e \epsilon)$ . Using the fact that for an  $n$ -dimensional vector  $X = [X_1, \dots, X_n]^T$ ,  $h(X) = h(X_1) + \sum_{i=2}^n h(X_i | X_{[1,i-1]}) \leq \sum_{i=1}^n h(X_i)$ , it follows that with  $E[\|x_t - \hat{x}_t\|^2] \leq \epsilon_t$ ,  $h(x_t - \hat{x}_t) \leq \frac{n}{2} \log(2\pi e \epsilon_t)$ . The final equality then follows from the fact that conditioned on  $x_{t-1}$ ,  $x_t$  and  $q'_{[0,t-1]}$  are independent. For the final result, in (14), taking the limit as  $\epsilon \rightarrow 0$ ,  $\log(\epsilon) \rightarrow -\infty$ , and  $C_0 = \infty$  follows.  $\square$

We note that the result also applies to the case when  $w_t$  is not i.i.d., but is stationary. Here, one needs to consider  $\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{[0,t-1]}) > 0$ , however.

The proof presented above in part builds on what is known as Shannon's Lower Bounding technique; this method is commonly used in information theory (see e.g. [26]).

We now consider the asymptotic almost sure criterion (6) for additive noise systems on  $\mathbb{R}^N$ .

**Theorem 4.2** *Suppose that  $X = W = \mathbb{R}^N$  and the system is given by  $x_{t+1} = f(x_t) + w_t$ , where the noise measure  $\nu$  admits a bounded density which is positive everywhere. Then, under (6),  $C_0 = \infty$ .*

**Proof:** For a given time stage  $t > 0$ , let  $q'_0, \dots, q'_{t-1}$  be given. Due to the additive nature of the noise, for all Borel sets  $B$ , it follows that for some  $K \in \mathbb{R}_+$ ,

$$P(x_t \in B | x_{t-1}) = \nu(B - f(x_{t-1})) \leq K \lambda(B),$$

implying that this measure is bounded from above by a constant multiple of the Lebesgue measure. Given a communication rate for a given time stage  $t$ , under any encoding and decoding policy with the finite partitioning of the state space  $X$  for encoding  $x_t$  leading to  $\hat{x}_t$ , there exists  $\bar{\epsilon}$  so that for all  $\epsilon < \bar{\epsilon}$ , the set

$$A_k(q'_0, \dots, q'_t) = \left\{ x_t : d(x_t, \hat{x}_t(q'_0, \dots, q'_{t-1}, q'_t)) \geq \epsilon \right\},$$

has a positive measure bounded from below by (since  $\mathbb{R}^N \setminus A_k(q'_0, \dots, q'_t)$  is a bounded set)

$$\begin{aligned} 1 - \sum_{q' \in \mathcal{M}'} P\left(x_t \notin \left\{ x : d(x, \hat{x}_t(q'_0, q'_1, \dots, q'_{t-1}, q'_t = q')) \leq \epsilon \right\} \middle| x_{[0,t-1]}, q'_{[0,t-1]}\right) \\ \geq 1 - |\mathcal{M}'| K \lambda(B_\epsilon) > 0, \end{aligned}$$

where  $\lambda(B_\epsilon)$  is the Lebesgue measure of the ball  $B_\epsilon$ .

This implies that, uniform over  $q'_0, \dots, q'_{t-1}$

$$P\left(x_t \in A_k(q'_0, \dots, q'_{t-1}) \middle| x_{t-1}, x_{t-1} \notin A_k(q'_0, \dots, q'_{t-1})\right) > 0 \quad (15)$$

uniform over all realizations of  $x_{t-1} \in A_k(q'_0, \dots, q'_{t-1})$ .

With

$$C_t = \left\{ x_t : d(x_t(x_0, w_0, \dots, w_{t-2}, w_{t-1}), \hat{x}_t(q'_0, \dots, q'_{t-1}, q'_t)) \geq \epsilon \right\}$$

and

$$C'_t = \left\{ x_t : d(x_t(x_0, w_0, \dots, w_{t-2}, w_{t-1}), \hat{x}_t(q'_0, \dots, q'_{t-1}, q'_t)) < \epsilon \right\}$$

and

$$B_t := \left\{ w_t \in \cup_{x_{t-1} \notin A_k(q'_0, \dots, q'_{t-1})} A_k(q'_0, \dots, q'_t) - f(x_{t-1}) \right\},$$

let  $\eta_C := \sum_{k=1}^{\infty} 1_{C_k}$  and  $\eta_{C'} := \sum_{k=1}^{\infty} 1_{C'_k}$ , where  $1_E$  stands for the indicator function of the event  $E$ .

Our goal is to show that  $\eta_C = \infty$  almost surely, leading to the desired conclusion. It is evident that  $\eta_C = \infty$  or  $\eta_{C'} = \infty$  almost surely. Suppose that  $\eta_{C'} = \infty$ . We will show that this implies that  $\eta_C$  is infinite as well. Note now that the event  $C_t$  is implied by the simultaneous occurrence of the events  $C'_{t-1}$  and  $B_t$ . Let

$$\tau_C(1) = \min(k > 0 : 1_{C_k} = 1) \text{ and } \tau_C(z) = \min(k > \tau_C(z-1) : 1_{C_k} = 1).$$

It is evident that  $P(\tau_C(1) < \infty) = 1$  by a repeated use of (15), since the event  $\tau_C(1) = \infty$  would imply an infinite number of violating the event whose probability is lower bounded by (15). Thus,  $P(\eta_C \geq 1) = 1$ . Since  $w_t$  is i.i.d., by a repeated use of (15) and induction,

$$P(\eta_C \geq k) = P(\eta_C \geq k, \eta_C \geq k-1) = P(\tau_C(1) < \infty | \mathcal{F}_{\tau_C(k-1)}) P(\eta_C \geq k-1) = 1,$$

where  $\mathcal{F}_{\tau_C(k-1)}$  is the  $\sigma$ -field generated by  $\{x_s, q'_s\}$  up to time  $\tau_C(k-1)$  and the last inequality follows since  $P(\tau_C(1) < \infty | \mathcal{F}_{\tau_C(k-1)}) = 1$  by the strong Markov property. Thus, for every  $k \in \mathbb{N}$ ,  $P(\eta_C \geq k) = 1$ , and it follows by continuity in probability that  $P(\eta_C = \infty) = \lim_{k \rightarrow \infty} P(\eta_C \geq k) = 1$ . Hence, for any finite communication rate, almost sure boundedness is not possible for arbitrarily small  $\epsilon > 0$ .  $\square$

Now let  $x_t$  be a stationary source and suppose that the goal is to achieve a uniform lower bound on the estimation error in the quadratic sense.

We first revisit and state a general and a quite strong result due to Graf and Luschgy [13] (further relevant results are due to various authors [52, 38, 12, 17]) on the asymptotic performance of optimal quantizers.

**Theorem 4.3** *Let  $X$  be an  $\mathbb{R}^k$ -valued random variable with density  $p$ . Let  $\mathcal{Q}_n := \{Q_n : \mathbb{R}^N \rightarrow \mathcal{M}, |\mathcal{M}| \leq n\}$  denote the set of all quantizers with cardinality at most  $n$ . Let*

$$V_n(X) = \inf_{Q \in \mathcal{Q}_n, \gamma^d : \mathcal{M} \rightarrow \mathbb{R}^N} E \left[ \|X - \gamma^d(Q_n(X))\|^2 \right].$$

Then,

(a)

$$\liminf_{n \rightarrow \infty} n^{\frac{r}{k}} V_n(X) \geq K_2 \|p\|_{\frac{k}{k+2}},$$

where  $\inf_{n \geq 1} n^{(2/k)} M_{n,2}([0, 1]^k) = K_2$  with

$$M_{n,2}([0, 1]^k) = \frac{V_n(U[0, 1]^k)}{(\lambda^k([0, 1]^k))^{(2/d)}}$$

is a coefficient for optimal quantization of the uniformly distributed source on  $[0, 1]^k$  and

$$\|p\|_{\frac{k}{k+2}} = \left( \int p^{\frac{k}{k+2}}(x) dx \right)^{\frac{k+2}{k}}.$$

(b) If  $E[\|X\|_{2+\delta}] < \infty$  for some  $\delta > 0$ , then

$$\lim_{n \rightarrow \infty} n^{\frac{2}{k}} V_n(X) = K_2 \left( \int p^{\frac{k}{k+2}}(x) dx \right)^{\frac{k+2}{k}},$$

For scalar sources, i.e. such with  $k = 1$ , we note that the value  $K_2$  is equal to  $\frac{1}{12}$ , cf. [12, 17].

An implication of the result above is the following.

**Theorem 4.4** Assume that the stationary source is such that  $x_t$  admits a density and that  $E[\|x_t\|_{2+\delta}] < \infty$  for some  $\delta > 0$ . Suppose further that the channel under consideration is a discrete noiseless channel. Under (7), memoryless quantization of the marginal source process leads to a distortion upper bound for any  $t$  given by the following so that for every  $\eta > 0$ , there exists  $\bar{\epsilon}$  so that for all  $\epsilon_t \leq \bar{\epsilon}$ , the following holds:

$$\epsilon_t := E[\|x_t - \hat{x}_t\|^2] \leq K_2(L_C + \eta)2^{-2C/N},$$

where

$$\lim_{C \rightarrow \infty} L_C = \|p\|_{\frac{N}{N+2}}$$

with  $C$  being the channel capacity. In particular, for (7) we have, for every  $\eta > 0$  and for sufficiently small  $\epsilon$ ,

$$C_\epsilon \leq \frac{N}{2} \log\left(\frac{K_2(\|p\|_{\frac{N}{N+2}} + \eta)}{\epsilon}\right).$$

**Proof:** The proof follows by encoding  $x_t$  in a memoryless fashion, letting  $n = 2^C$ , and applying Theorem 4.3.  $\square$

One could also obtain a lower bound for the special case of additive systems, by encoding only the noise process.

**Theorem 4.5** Assume that the system is given by

$$x_{t+1} = f(x_t) + w_t,$$

and  $w_t \sim \nu$ , where  $\nu$  admits a density  $p_w$ . Further assume that the channel is of Class A type. Then

$$C_\epsilon < \frac{N}{2} \log\left(\frac{K_2\|p_w\|_{\frac{N}{N+2}}}{\epsilon}\right)$$

implies that there exists no encoder and decoder which can achieve (7) for  $\epsilon$  sufficiently small. In particular,  $C_0 = \infty$ .

**Proof:** The proof follows from the fact that the estimation error for the noise process gives a lower bound for the estimation error of the dynamical system itself. In particular, given any encoding policy, the following holds:

$$\begin{aligned} & \inf_{\hat{x}_t = \gamma_t(q'_{[0,t]})} E[\|x_t - \hat{x}_t\|^2 | q'_{[0,t]}] \\ & \geq \inf_{\hat{x}_t = \gamma_t(q'_{[0,t]}, x_{t-1})} E[\|x_t - \hat{x}_t\|^2 | q'_{[0,t]}, x_{t-1}] \\ & = \inf_{\hat{x}_t = \gamma_t(q'_{[0,t]}, x_{t-1})} E[\|f(x_{t-1}) + w_{t-1} - \hat{x}_t\|^2 | q'_{[0,t]}, x_{t-1}] \\ & = \inf_{\hat{x}_t = \gamma_t(q'_{[0,t]}, x_{t-1})} E[\|w_{t-1} - (\hat{x}_t - f(x_{t-1}))\|^2 | q'_{[0,t]}, x_{t-1}] \\ & = \inf_{\hat{v}_t = \gamma_t(q'_{[0,t]}, x_{t-1})} E[\|w_{t-1} - \hat{v}_t\|^2 | q'_{[0,t]}, x_{t-1}] \end{aligned}$$

Thus, a lower bound is obtained by trying to estimate the noise realizations, given the additional side information of the previous state realization. At any given time the conditional probability  $P(dw_t | q'_{[0,t-1]}, x_{t-1}) = P(dw_t)$  can be partitioned into at most  $|\mathcal{M}|$  bins, leading to the applicability of the asymptotic quantization theory results for each noise realization. The argument then follows from Theorem 4.3(a).  $\square$



## 5 A random dynamical systems view

In this section, the system (4) is viewed as a random dynamical system (briefly an RDS). We will prove that under the assumption of a compact state space and a noiseless channel, the metric entropy of the RDS is a lower bound on  $C_0$  for the objective (6), and hence also for (5).

Next we define a random dynamical system. The base space is defined as  $B := W^{\mathbb{Z}_+}$  and for its elements we use the notation  $\bar{w} = (w_t)_{t \geq 0}$ . The  $\sigma$ -field on  $B$  is the product field of the Borel  $\sigma$ -field on  $W$  (generated by cylinder sets), and the measure on this space is  $\nu^{\mathbb{Z}_+}$ , the corresponding product measure of  $\nu$ . The dynamics on  $B$  is given by the left shift operator  $(\theta \bar{w})_t = w_{t+1}$ , which obviously preserves the measure  $\nu^{\mathbb{Z}_+}$  and is easily seen to be ergodic. A random dynamical system (briefly an *RDS*) over  $\theta$  is given by

$$\varphi(t, \bar{w})x := f_{w_{t-1}} \circ \cdots \circ f_{w_1} \circ f_{w_0}(x), \quad \varphi : \mathbb{Z}_+ \times B \rightarrow \text{Homeo}(X).$$

The associated skew-product transformation is given by

$$\Phi : \mathbb{Z}_+ \times B \times X \rightarrow B \times X, \quad (t, (\bar{w}, x)) \mapsto \Phi_t(\bar{w}, x) = (\theta^t \bar{w}, \varphi(t, \bar{w})x).$$

We will work with a time-invertible extension of  $\varphi$ , which replaces  $B$  with  $B^* := W^{\mathbb{Z}}$  (two-sided sequences) endowed with the measure  $\nu^{\mathbb{Z}}$  and the shift operator  $\theta^*$  on two-sided sequences. Since  $f_w$  is invertible by assumption for every  $w \in W$ , it is obvious that also the cocycle  $\varphi$  over  $(B, \theta)$  can be extended to a cocycle

$$\varphi^* : \mathbb{Z} \times B^* \times X \rightarrow X$$

over  $(B^*, \theta^*)$ . For simplicity, we drop the superscript  $*$  again and just write  $(\theta, \varphi)$  for the time-invertible extension.

For technical reasons, we also extend the one-sided sequence  $(w_t)_{t \geq 0}$  of random variables by a two-sided sequence  $(w_t)_{t \in \mathbb{Z}}$ , which is still i.i.d. such that  $w_t \sim \nu$ .

Consider a probability measure  $\mu$  on  $B \times M$ , invariant under the time-1-map  $\Phi_1 : B \times M \rightarrow B \times M$ , with marginal  $\nu^{\mathbb{Z}}$  on  $B$ . Then  $\mu$  is called a  $\varphi$ -invariant measure. Such  $\mu$  disintegrates with respect to  $\nu^{\mathbb{Z}}$ , i.e.,  $d\mu(\bar{w}, x) = d\mu_{\bar{w}}(x) d\nu^{\mathbb{Z}}(\bar{w})$  for a family  $\{\mu_{\bar{w}}\}$  of conditional probabilities, defined  $\nu^{\mathbb{Z}}$ -almost everywhere.

The general definition of the metric entropy  $h_\mu(\varphi)$  of an RDS  $\varphi$  with respect to an invariant measure  $\mu$  can be found in [3]. We will use a characterization of the metric entropy in terms of so-called  $(\bar{w}, T, \epsilon)$ -spanning sets, which generalizes Katok's characterization for the deterministic case (see Subsect. 2). First, define for every  $\bar{w} \in B$  the family of Bowen-metrics

$$d_{\bar{w}}^T(x, y) := \max_{0 \leq t \leq T-1} d(\varphi(t, \bar{w})x, \varphi(t, \bar{w})y), \quad T \geq 1.$$

We say that a set  $F \subset M$   $(\bar{w}, T, \epsilon)$ -spans another set  $A \subset M$  if for each  $x \in A$  there is  $y \in F$  with  $d_{\bar{w}}^T(x, y) \leq \epsilon$ . A set  $E \subset M$  is  $(\bar{w}, T, \epsilon)$ -separated if for each  $x, y \in E$  with  $x \neq y$  it holds that  $d_{\bar{w}}^T(x, y) > \epsilon$ . By  $s_A(\bar{w}, T, \epsilon)$  we denote the smallest cardinality a  $(\bar{w}, T, \epsilon)$ -spanning set for  $A$  can have. We use the notation  $r_A(\bar{w}, T, \epsilon)$  to denote the maximal cardinality of a  $(\bar{w}, T, \epsilon)$ -separated set contained in some  $A \subset M$ .

Now let  $\mu$  be an ergodic invariant measure for the RDS  $\varphi$  (i.e.,  $\mu$  is ergodic w.r.t. the skew-product  $\Phi$ ). Then  $\mu$  disintegrates as  $d\mu(\bar{w}, x) = d\mu_{\bar{w}}(x) d\nu^{\mathbb{Z}}(\bar{w})$  with almost everywhere defined sample measures  $\mu_{\bar{w}}$ . For any  $\delta \in (0, 1)$  put

$$s(\bar{w}, T, \epsilon, \delta) := \min \{s_A(\bar{w}, T, \epsilon) : A \in \mathcal{B}(M), \mu_{\bar{w}}(A) \geq 1 - \delta\}.$$

We will use the following generalization of Katok's result from [54, Thm. 3.1], which describes the metric entropy of an RDS for an ergodic measure via cardinalities of  $(\bar{w}, T, \epsilon)$ -spanning sets.

**Theorem 5.1** *If  $h_\mu(\varphi) < \infty$ , then for any  $\delta \in (0, 1)$ ,*

$$h_\mu(\varphi) = \lim_{\epsilon \downarrow 0} \limsup_{T \rightarrow \infty} \frac{1}{T} \log s(\bar{w}, T, \epsilon, \delta) \quad \nu^{\mathbb{Z}}\text{-a.s.}$$

In addition to this theorem, we will make use of the following technical lemma.

**Lemma 5.1** *Let  $(\Omega, \mathcal{F}, P)$  be a probability space,  $f : \Omega \rightarrow [0, 1]$  an integrable function and  $p \in (0, 1)$  such that*

$$\int_{\Omega} f(\omega) dP(\omega) \geq p.$$

*Then  $f(\omega) \geq p/2$  for all  $\omega$  in a set of measure  $\geq p/(2-p)$ .*

**Proof:** Assume to the contrary that

$$P\left(\left\{\omega \in \Omega : f(\omega) \geq \frac{p}{2}\right\}\right) =: \alpha < \frac{p}{2-p}.$$

This is equivalent to

$$P\left(\left\{\omega \in \Omega : f(\omega) < \frac{p}{2}\right\}\right) = 1 - \alpha > 2\frac{1-p}{2-p}.$$

Splitting  $\Omega$  into the set  $\Omega_1$ , where  $f(\omega) \geq p/2$ , and its complement  $\Omega_2$ , we see that

$$\begin{aligned} p &\leq \int_{\Omega_1} f(\omega) dP(\omega) + \int_{\Omega_2} f(\omega) dP(\omega) \leq P(\Omega_1) + \frac{p}{2}P(\Omega_2) \\ &= \alpha + \frac{p}{2}(1 - \alpha) = \alpha \left(1 - \frac{p}{2}\right) + \frac{p}{2} < \frac{p}{2-p} \frac{2-p}{2} + \frac{p}{2} = p, \end{aligned}$$

a contradiction. □

Now we are in position to prove the announced theorem. In the proof we will also use the following concepts and notations: For any  $T > T_0$ , we say that a set  $E \subset M$  is  $(\bar{w}, T, \epsilon; T_0)$ -separated if  $d(\varphi(t, \bar{w})x, \varphi(t, \bar{w})y) > \epsilon$  for some  $t \in \{T_0, T_0 + 1, \dots, T - 1\}$ , whenever  $x, y \in E$  and  $x \neq y$ . We write  $r(\bar{w}, T, \epsilon; T_0)$  for the maximal cardinality of a  $(\bar{w}, T, \epsilon; T_0)$ -separated set.

**Theorem 5.2** *Consider the estimation objective (6) to be achieved for an initial measure  $\pi_0$  which is equivalent to an ergodic measure  $\pi$  of the Markov chain  $\{x_t\}_{t \geq 0}$ . Then there exists an ergodic measure  $\mu$  for the random dynamical system  $\varphi$  such that*

$$C_0 \geq h_{\mu}(\varphi)$$

*holds if  $h_{\mu}(\varphi) < \infty$ .*

**Proof:** In the proof of Lemma 3.1 we showed that the product measure  $\nu^{\mathbb{Z}^+} \times \pi$  is invariant under the RDS with one-sided time. It is well-known that this measure has a unique extension to an ergodic invariant measure  $\mu$  for the two-sided system (cf. [21]). We write  $d\mu(\bar{w}, x) = d\mu_{\bar{w}}(x) d\nu^{\mathbb{Z}}(\bar{w})$  for the disintegration of  $\mu$ . Now we prove the theorem in three steps.

*Step 1.* Consider a noiseless channel with input alphabet  $\mathcal{M}$  and a pair of coder and decoder/estimator which solves the estimation problem (6) for some  $\epsilon > 0$ . For all  $(\bar{w}, x) \in B \times M$  and  $\delta > \epsilon$  let

$$T^{\bar{w}}(x, \delta) := \inf \left\{ k \in \mathbb{N} : \sup_{t \geq k} d(\varphi(t, \bar{w})x, \hat{x}_t) \leq \delta \right\},$$

where the infimum is defined as  $+\infty$  if the corresponding set is empty. Note that  $T^{\bar{w}}(x, \delta)$  depends measurably on  $(\bar{w}, x)$ , since  $(\bar{w}, x) \mapsto \varphi(t, \bar{w})x$  is measurable and also the coder and estimator maps are assumed to be measurable. Define

$$B^K(\delta) := \{(\bar{w}, x) \in B \times M : T^{\bar{w}}(x, \delta) \leq K\} \quad \text{for all } \delta > \epsilon \text{ and } K \in \mathbb{N},$$

and observe that these sets are measurable. For every  $\bar{w} \in B$  we write

$$A^K(\bar{w}, \delta) := \{x \in M : (\bar{w}, x) \in B^K(\delta)\}.$$

From (6) it follows that for every  $\delta > \epsilon$ ,

$$\lim_{K \rightarrow \infty} P(B^K(\delta)) = P\left(\bigcup_{K \in \mathbb{N}} B^K(\delta)\right) = 1.$$

Here the measure  $P$  is the two-sided extension of the product measure  $\nu^{\mathbb{Z}+} \times \pi_0$ . Since we assume that  $\pi$  is absolutely continuous with respect to  $\pi_0$ , for every  $\alpha > 0$  there exists  $\beta > 0$  so that  $(\nu^{\mathbb{Z}+} \times \pi)(C) < \alpha$  if  $(\nu^{\mathbb{Z}+} \times \pi_0)(C) < \beta$ . This property carries over to the two-sided extensions of these product measures. Applying this to the complements of the sets  $B^K(\delta)$ , we find that

$$\lim_{K \rightarrow \infty} \int_B \mu_{\bar{w}}(A^K(\bar{w}, \delta)) d\nu^{\mathbb{Z}}(\bar{w}) = 1. \quad (16)$$

Consider the sequence  $\epsilon_n := 1/n$ ,  $n \in \mathbb{N}$ . For each fixed  $n$  we can choose  $K = K(n)$  large enough so that

$$\int_B \mu_{\bar{w}}(A^{K(n)}(\bar{w}, 2\epsilon_n)) d\nu^{\mathbb{Z}}(\bar{w}) \geq 2 \frac{1 - 3^{-n}}{2 - 3^{-n}} =: p_n. \quad (17)$$

This is possible, since  $p_n \in (0, 1)$  for all  $n \in \mathbb{N}$ . Observe that

$$\lim_{n \rightarrow \infty} \frac{p_n}{2 - p_n} = \lim_{n \rightarrow \infty} (1 - 3^{-n}) = 1.$$

By Lemma 5.1 we find a set  $\tilde{B} = \tilde{B}(n) \subset B$  of measure

$$\nu^{\mathbb{Z}}(\tilde{B}(n)) \geq 1 - 3^{-n} \quad (18)$$

so that for all  $\bar{w} \in \tilde{B}(n)$  we have

$$\mu_{\bar{w}}(A^{K(n)}(\bar{w}, 2\epsilon_n)) \geq \frac{1 - 3^{-n}}{2 - 3^{-n}} \geq \frac{1}{4}. \quad (19)$$

*Step 2.* In this step, we fix  $n$  and write  $\delta = 2\epsilon_n$ ,  $K = K(n)$ ,  $\tilde{B} = \tilde{B}(n)$ . For a given channel and coder-decoder pair which solves the estimation problem for  $\epsilon = \epsilon_n$  define for every integer  $T > 0$  the set

$$\hat{E}(T) := \{\hat{x}_{[0, T-1]} : q_{[0, T-1]} \in \mathcal{M}^T\},$$

i.e., the set of all possible estimation sequences of length  $T$ . Let  $E \subset A^K(\bar{w}, \delta)$  be  $(\bar{w}, K + T, 2\delta; K)$ -separated for a fixed  $\bar{w} \in \tilde{B}$  and some  $T \in \mathbb{N}$ . Define a map  $\alpha : E \rightarrow \hat{E}(T)$ , by assigning to each  $x \in E$  an  $\hat{x}_{[0, T-1]} \in \hat{E}(T)$  with

$$d(\varphi(t, \bar{w})x, \hat{x}_t) \leq \delta \quad \text{for } K \leq t \leq K + T - 1,$$

which is possible by the definition of  $A^K(\bar{w}, \delta)$ . The map  $\alpha$  is injective, because  $\alpha(x) = \alpha(y)$  by the triangle inequality implies

$$d(\varphi(t, \bar{w})x, \varphi(t, \bar{w})y) \leq 2\delta, \quad K \leq t \leq K + T - 1,$$

and hence  $x = y$ , since  $E$  is  $(\bar{w}, K + T, 2\delta; K)$ -separated. Consequently,  $|E| \leq |\hat{E}(T)| \leq |\mathcal{M}|^T$ , implying

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \log r_{A^K(\bar{w}, \delta)}(\bar{w}, K + T, 2\delta; K) \leq \log |\mathcal{M}|.$$

If  $E \subset \varphi(K, \bar{w})A^K(\bar{w}, \delta)$  is a  $(\theta^K \bar{w}, T, 2\delta)$ -separated set, then  $E' := \varphi(K, \bar{w})^{-1}(E)$  is a  $(\bar{w}, K + T, 2\delta; K)$ -separated set of the same cardinality, implying

$$r_{\varphi(K, \bar{w})A^K(\bar{w}, \delta)}(\theta^K \bar{w}, T, 2\delta) \leq r_{A^K(\bar{w}, \delta)}(\bar{w}, K + T, 2\delta; K),$$

and with  $C(\bar{w}) := \varphi(K, \bar{w})A^K(\bar{w}, \delta)$  we thus obtain

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \log r_{C(\bar{w})}(\theta^K \bar{w}, T, 2\delta) \leq \log |\mathcal{M}|.$$

Using the invariance property of the sample measures (see [21, Prop. 1.2.3]) and (19), we find

$$\mu_{\Theta^K \bar{w}}(C(\bar{w})) = (\varphi(K, \bar{w})\mu_{\bar{w}})(C(\bar{w})) = \mu_{\bar{w}}(A^K(\bar{w}, \delta)) \geq \frac{1}{4}.$$

Since the channel was chosen arbitrarily (within the class of channels that allow for the estimation criterion with accuracy  $\epsilon$ ), we have

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \log r_{C(\theta^{-K(n)} \bar{w})}(\bar{w}, T, 4\epsilon_n) \leq C_{\epsilon_n} \quad \text{for all } \bar{w} \in \theta^{K(n)}(\tilde{B}(n)). \quad (20)$$

*Step 3.* We claim that the set  $B'$  defined as follows has positive measure:

$$B' := \bigcap_{n=1}^{\infty} \theta^{K(n)}(\tilde{B}(n)).$$

Indeed, using (18) and the fact that  $\nu^{\mathbb{Z}}$  is invariant under  $\theta$ , we find

$$\begin{aligned} \nu^{\mathbb{Z}}(B') &= 1 - \nu^{\mathbb{Z}}\left(\bigcup_{n=1}^{\infty} [\theta^{K(n)}(\tilde{B}(n))]^c\right) \geq 1 - \sum_{n=1}^{\infty} \nu^{\mathbb{Z}}([\theta^{K(n)}(\tilde{B}(n))]^c) \\ &= 1 - \sum_{n=1}^{\infty} (1 - \nu^{\mathbb{Z}}(\tilde{B}(n))) \geq 1 - \sum_{n=1}^{\infty} 3^{-n} = \frac{1}{2}. \end{aligned}$$

Hence, there are generic elements  $\bar{w} \in B'$  (i.e., elements that are contained in an arbitrary but fixed full measure subset of  $B$ ) so that (with (20))

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \log r_{A_n}(\bar{w}, T, 4\epsilon_n) \leq C_{\epsilon_n} \quad \text{for all } n \in \mathbb{N},$$

where  $A_n$  is a set with  $\mu_{\bar{w}}(A_n) \geq 1/4$ . Fixing such a generic element  $\bar{w}$  with respect to the full measure set provided by Theorem 5.1 with  $\delta = 1/4$ , we find

$$h_{\mu}(\varphi) \leq \lim_{n \rightarrow \infty} \limsup_{T \rightarrow \infty} \frac{1}{T} \log r_{A_n}(\bar{w}, T, 4\epsilon_n) \leq \lim_{n \rightarrow \infty} C_{\epsilon_n} = C_0.$$

Here we use the simple fact that a maximal  $(\bar{w}, T, 4\epsilon_n)$ -separated set in  $A_n$  also  $(\bar{w}, T, 4\epsilon_n)$ -spans  $A_n$  and the assumption that  $h_{\mu}(\varphi) < \infty$ .  $\square$

**Remark 5.1** *It should be mentioned that the metric entropy  $h_{\mu}(\varphi)$  in the theorem can be expressed in terms of the positive Lyapunov exponents of the RDS  $\varphi$ , given that the state space is a smooth manifold and the system satisfies some regularity assumptions. In particular, this includes that  $f$  is twice differentiable with respect to  $x$  and the invariant measure  $\mu$  satisfies the so-called SRB property, cf. [21].*

## 6 Almost sure estimation for noise-free systems

In this section, we study the estimation objective (6) in the case when there is only noise in the channel, but not in the source.

## 6.1 A general result for non-linear systems

The following result shows that in case of a noise-free system, the topological entropy provides an upper bound on  $C_0$  for the objective (6). The proof uses similar arguments as employed in [29] for the case of a noiseless channel.

**Theorem 6.1** *Consider a non-linear deterministic system  $x_{t+1} = f(x_t)$  on a compact state space  $X$ , estimated via a discrete memoryless channel (DMC). Then, for the asymptotic estimation objective (6),*

$$C_0 \leq h_{\text{top}}(f).$$

**Proof:** Without loss of generality, we may assume that  $h_{\text{top}}(f) < \infty$ , since otherwise the statement trivially holds. Then it suffices to show that for any  $\epsilon > 0$  the estimation objective can be achieved whenever the channel capacity satisfies  $C > h_{\text{top}}(f)$ . Since the capacity of a DMC can take any positive value, it follows that  $C_\epsilon \leq h_{\text{top}}(f)$  for every  $\epsilon > 0$  and thus  $C_0 \leq h_{\text{top}}(f)$ .

Now, consider a channel with capacity  $C > h_{\text{top}}(f)$  and fix  $\epsilon > 0$ . Recall that the input alphabet is denoted by  $\mathcal{M}$  and the output alphabet by  $\mathcal{M}'$ . By the random coding construction of Shannon [10], we can achieve a rate  $R$  satisfying

$$h_{\text{top}}(f) < R < C \quad (21)$$

with a sequence of increasing sets  $\{1, \dots, M_n\}$  of input messages so that for all  $n$ ,

$$2^{nR} \leq M_n \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{1}{n} \log M_n = C. \quad (22)$$

Furthermore, there exists a sequence of encoders  $E^n : \{1, \dots, M_n\} \rightarrow \mathcal{M}^n$ , yielding codewords  $x^n(1), \dots, x^n(M_n)$ , and a sequence of decoders  $D^n : (\mathcal{M}')^n \rightarrow \{1, \dots, M_n\}$  so that

$$P(D^n(q'_{[0,n-1]}) \neq c | q_{[0,n-1]} = x^n(c)) \leq e^{-nE(R)+o(n)},$$

uniformly for all  $c \in \{1, \dots, M_n\}$ . Here  $\frac{o(n)}{n} \rightarrow 0$  as  $n \rightarrow \infty$  and  $E(R) > 0$ . In particular, we observe that with  $c_n \in \{1, \dots, M_n\}$  being the message transmitted and  $D^n(q'_{[0,n-1]})$  the decoder output,

$$\begin{aligned} & P(D^n(q'_{[0,n-1]}) \neq c_n) \\ &= \sum_{c \in \{1, \dots, M_n\}} P(D^n(q'_{[0,n-1]}) \neq c | q_{[0,n-1]} = x^n(c)) P(q_{[0,n-1]} = x^n(c)) \leq e^{-nE(R)+o(n)}. \end{aligned}$$

This also implies that the bound holds even when the messages to be transmitted are not uniformly distributed. Thus, for the sequence of encoders and decoders constructed above we have

$$\sum_n P(D^n(q'_{[0,n-1]}) \neq c_n) \leq \sum_n e^{-nE(R)+o(n)} < \infty.$$

The Borel-Cantelli Lemma then implies

$$P\left(\left\{D^n(q'_{[0,n-1]}) \neq c_n \text{ infinitely often}\right\}\right) = 0. \quad (23)$$

Now we choose  $\delta \in (0, \epsilon)$  so that, by uniform continuity,

$$d(x, y) < \delta \quad \Rightarrow \quad d(f(x), f(y)) < \epsilon \quad \text{for all } x, y \in X. \quad (24)$$

Furthermore, we choose  $N$  sufficiently large so that

$$r_{\text{span}}(n, \delta) \leq M_n \quad \text{for all } n \geq N. \quad (25)$$

This is possible, because by (21) and (22), for every  $n \in \mathbb{N}$  we have

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log r_{\text{span}}(n, \delta) \leq h_{\text{top}}(f) < R = \frac{1}{n} \log 2^{nR} \leq \frac{1}{n} \log M_n.$$

Let  $S_j$  be a  $(j, \epsilon)$ -spanning set of cardinality  $r_{\text{span}}(j, \delta)$  and fix injective functions

$$\iota_j : S_j \rightarrow \{1, \dots, M_j\}.$$

In fact, by possibly enlarging the set  $S_j$ , we can assume that  $\iota_j$  is bijective. For any  $a \in X$  let  $x_j^*(a)$  denote a fixed element of  $S_j$  satisfying  $d(f^t(x_j^*(a)), f^t(a)) \leq \delta$  for  $0 \leq t \leq j - 1$ .

Define sampling times by

$$\tau_0 := 0 \quad \text{and} \quad \tau_{j+1} := \tau_j + j + 1 \text{ for } j \geq 0.$$

In the following, we specify the coding scheme. In this coding scheme, the encoder from  $\tau_j$  to  $\tau_{j+1} - 1$ , encodes the information regarding the orbit of the state from  $\tau_{j+1}$  to  $\tau_{j+2} - 1$ . For all  $j \geq N$ , at time  $\tau_j$ , use the input  $\iota_{j+1}(x_{j+1}^*(f^{j+1}(x_{\tau_j})))$  for the encoder, where  $x_{\tau_j}$  is the state at time  $\tau_j$ . Then  $x^{j+1}(\iota_{j+1}(x_{j+1}^*(f^{j+1}(x_{\tau_j}))))$  is sent during the next  $j + 1$  units of time. This is possible by (25). For  $j < N$ , it is not important what we transmit.

Let the estimator apply  $x_{j+1}^* \circ \iota_{j+1}^{-1}$  to the output of the decoder, obtaining an element  $y_{j+1} \in S_{j+1}$ , and use  $y_{j+1}, f(y_{j+1}), \dots, f^j(y_{j+1}), f^{j+1}(y_{j+1})$  as the estimates during the forthcoming time interval of length  $\tau_{j+2} - \tau_{j+1} = \tau_{j+1} - \tau_j + 1 = j + 2$ . Then  $\delta < \epsilon$ , (24) and the fact that  $S_{j+1}$  is  $(j + 1, \delta)$ -spanning implies that the desired estimation accuracy is achieved, provided that there was no error in the transmission.

Now (23) implies that after a finite random time, there are no more errors in the transmission. By the analysis above, the errors will be uniformly bounded by  $\epsilon$ . Hence, the objective (6) is achieved.  $\square$

We note that the proof above crucially depends on the fact that the system is deterministic. Note also that the proof above admittedly is impractical to implement. The theorem is essentially a possibility result. Note that the proof even does not make use of the fact that the encoder has access to the realizations of the channel output, hence feedback is not utilized.

In the following, for noiseless linear systems, we will show that there exist stationary policies which guarantee the almost sure estimation criterion.

## 6.2 Refinement for linear systems

We consider a noiseless linear system

$$x_{t+1} = Ax_t \tag{26}$$

with  $x_t \in \mathbb{R}^N$ . The following result gives a positive answer to the question whether the estimation objective (6) can be achieved, when no noise is present in the system, via an erasure channel of finite capacity. Recall that such a channel has  $\mathcal{M} = \{0, 1\}$  and  $\mathcal{M}' = \{0, 1, e\}$ , with

$$P(q'|q) = (1 - p)1_{\{q'=q\}} + p1_{\{q'=e\}},$$

where  $e$  denotes the erasure symbol and  $p$  is the erasure probability.

**Theorem 6.2** *Consider system (26) estimated over a memoryless erasure channel with finite capacity. Then, for (6), we have*

$$C_0 \leq \sum_{|\lambda_i| > 1} \log(|\lambda_i|). \tag{27}$$

**Remark 6.1** *We note that if one samples the system with a sufficiently large period and work with the sampled system, the upper bound in (27), normalized by the sampling period, can be made arbitrarily close to  $\sum_{|\lambda_i| > 1} \log(|\lambda_i|)$  which is the topological entropy of the linear system (26).*

**Proof:** We will show that the capacity bound is almost achievable. We first assume that the system is diagonalizable. Consider the eigenvalues  $\lambda_i > 1$ . Similar to the construction in [51] (see also [30] and [32] for related discussions with different constructions) where a uniform quantizer is employed; we can show that with  $r > 0$  and  $\Delta_t = E[\sup_{x_t, \hat{x}_t} |x_t - \hat{x}_t|^r]$ , a dynamical stochastic system can be obtained where for some  $\delta > 0$ ,

$$\Delta_{t+1}^i \leq \left( p(|\lambda_i|^r + (1-p)\frac{(|\lambda_i|)^r}{2^{rR_i}}) \right) \Delta_t^i,$$

where  $\Delta_t^i = E[|x_t^i - \hat{x}_t^i|^r]$  and  $R_i$  is the rate allocated for the  $i$ th unstable mode.

Now, suppose that

$$\kappa_{i,r} := \left( p(|\lambda_i|^r + (1-p)\frac{(|\lambda_i|)^r}{2^{rR_i}}) \right) < 1.$$

It follows that  $\Delta_t^i \rightarrow 0$ . Furthermore, for every  $\zeta > 0$ , by Markov's inequality

$$P(|x_t^i - \hat{x}_t^i| \geq \zeta) \leq \Delta_0^i \frac{\kappa_{i,r}^t}{\zeta^r}$$

and since

$$\sum_t P(|x_t^i - \hat{x}_t^i| \geq \zeta) < \sum_t \Delta_0^i \frac{\kappa_{i,r}^t}{\zeta^r} < \infty,$$

Borel-Cantelli Lemma implies that  $|x_t^i - \hat{x}_t^i| \leq \zeta$  infinitely often. This applies for every  $\zeta > 0$ , and thus it follows that convergence in almost sure sense applies.

Now, we take a careful look at the expression

$$\left( p(|\lambda_i|^r + (1-p)\frac{(|\lambda_i|)^r}{2^{rR_i}}) \right) < 1. \quad (28)$$

The proof above shows that for every  $r > 0$ , this condition implies the almost sure convergence. The claim is that if  $R_i(1-p) > \log(|\lambda_i|)$ , there exists some  $r > 0$  so that the above holds. Observing that at  $r = 0$  (28) is an equality, we can take the derivative of this expression and show that the right derivative is negative (defined as the limit  $r \downarrow 0$ ). To check that we observe that the derivative writes as

$$\log(|\lambda_i|)|\lambda_i|^r p + (1-p) \log(|\lambda_i|2^{-R_i})(|\lambda_i|2^{-R_i})^r \quad (29)$$

Finally, one can check by writing that the condition  $R_i(1-p) > \log(|\lambda_i|)$  implies (by substituting  $2^{-R_i} < |\lambda_i|^{\frac{-1}{1-p}}$ ) that (29) is less than 0 for sufficiently small  $r > 0$ .

The above applies for each dimension. The capacity of the erasure channel is given by  $\sum R_i(1-p)$ , where  $R_i$  is the rate allocated to the  $i$ th component.

The same discussion also applies for the non-diagonalizable setting. In this case, one could apply the following alternative reasoning for each Jordan block in the system; these Jordan blocks are decoupled.

Let  $\Upsilon_t = \begin{bmatrix} \Upsilon_t^1 \\ \Upsilon_t^2 \\ \vdots \\ \Upsilon_t^N \end{bmatrix}$  where  $\Upsilon_t^i$  denote the support of the uncertainty size  $x_t^i - \hat{x}_t^i$ . It can be shown that the evolution of  $\Delta_t$  evolves as

$$\Upsilon_t = \prod_{k=0}^{t-1} p_k A_{p_k} \Upsilon_0,$$

where  $p_t = 0$  with probability  $p$  and  $p_t = 1$  with probability  $1-p$ , and  $A_0$  is a Jordan matrix with eigenvalue  $\lambda_i$ , and  $A_1$  is a Jordan matrix with eigenvalue  $|\lambda_i|2^{-R_i}$ .

One can thus view the process  $\Upsilon_t$  as the outcome of the product of random matrices. In general, the computation of the limit of the product of a sequence of random matrices, even in the i.i.d. case, is a

very challenging problem, but in this case one observes that the diagonal term of the matrix  $\prod_{k=0}^{t-1} p_t A_{p_t}$  is always the product of  $|\lambda_i|$  and  $|\lambda_i|2^{-R_i}$ , multiplied according to the number of erasures and successful transmissions across the channel. By the strong law of large numbers, it then follows that if

$$p \log(|\lambda_i|) + (1-p)(\log(2^{-R_i}|\lambda_i|)) < 0,$$

the product of the random matrices converges to pointwise zero almost surely. This condition is identical to the condition  $R_i(1-p) > \log(|\lambda_i|)$ . Thus, the result follows.  $\square$

Sahai and Mitter [37] show in Corollary 5.3 and Theorem 4.3 of their paper that for a discrete memoryless channel indeed it suffices that  $C > \sum_{|\lambda_i| \geq 1} \log(|\lambda_i|)$  for the existence of encoder and controller policies leading to stochastic stability. This result is in agreement with Theorems 6.1 and 6.2.

## 7 Conclusion

In this paper, we considered three estimation objectives for stochastic non-linear systems  $x_{t+1} = f(x_t, w_t)$  with i.i.d. noise  $(w_t)$ , assuming that the estimator receives state information via a noisy channel of finite capacity.

We studied three fundamentally different approaches to obtain lower bounds for  $C_0$ , the smallest channel capacity above which the estimation objective under consideration can be achieved for arbitrarily small errors. These approaches and the corresponding results can be summarized as follows:

- (1) For noiseless channels, assuming that the initial measure  $\pi_0$  is stationary, we proved that  $C_0$  is bounded below by either the topological or the metric entropy of a shift dynamical system on the space of trajectories (Theorems 3.1 and 3.2). We saw that these lower bounds are infinite when the noise substantially influences the dynamics so that the space of relevant trajectories becomes too large.
- (2) For systems on Euclidean space and noisy channels, we provided several information-theoretic conditions enforcing  $C_0 = \infty$ . In particular, Theorem 4.1 shows that  $C_0 = \infty$  for the quadratic stability objective, whenever

$$\limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T-1} h(x_t | x_{t-1}) > 0.$$

Since  $h(x_t | x_{t-1})$  is a measure for the uncertainty of  $x_t$  given  $x_{t-1}$ , this condition says that the noise on the long run (in average) influences the state process in a substantial way. Similarly to the results in Section 3, this means that the noise makes the space of relevant trajectories too large (or too complicated) to estimate the state with arbitrarily small error over a finite capacity channel.

- (3) In Section 5 we viewed the stochastic system  $x_{t+1} = f(x_t, w_t)$  as a random dynamical system, defined as a cocycle over the shift on the space of noise realizations. Assuming a compact state space and a noiseless channel, we proved that the metric entropy of the random dynamical system provides a lower bound on  $C_0$  for the objective (6). Since this lower bound is finite under mild assumptions and in case of a smooth system its value only depends on the average divergence rates of nearby trajectories (i.e., Lyapunov exponents), we see that the random dynamical systems view is not shedding too much light on the problem.
- (4) In Section 6 we assumed that the system is deterministic with a compact state space, but the channel is noisy. We proved that in this case  $C_0$  is bounded above by the topological entropy of the system for the asymptotic almost sure objective. Furthermore, we showed the analogous result, when the system is linear and the channel is an erasure channel.

It would be interesting to generalize the analysis in this paper to controlled non-linear systems, as well as to establish further connections between the ergodic theory of random dynamical systems and information theory.



## References

- [1] H. Asnani and T. Weissman. On real time coding with limited lookahead. *Information Theory, IEEE Transactions on*, 59(6):3582–3606, 2013.
- [2] T. Berger. Information rates of Wiener processes. *IEEE Transactions on Information Theory*, 16:134–139, 1970.
- [3] T. Bogenschütz. Entropy, pressure, and a variational principle for random dynamical systems. *Random Comput. Dynam.*, 1(1):99–116, 1992.
- [4] V. A. Boichenko, G. A. Leonov, V. Reitmann. Dimension theory for ordinary differential equations. Teubner-Texte zur Mathematik, 141. Teubner, Stuttgart, 2005.
- [5] V. S. Borkar, S. K. Mitter, and S. Tatikonda. Optimal sequential vector quantization of Markov sources. *SIAM J. Control and Optimization*, 40:135–148, 2001.
- [6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
- [7] N. Şen, F. Alajaji, and S. Yüksel. Feedback capacity of a class of symmetric finite-state Markov channels. *IEEE Transactions on Information Theory*, 56:4110 – 4122, July 2011.
- [8] R. Dabora and A. Goldsmith. On the capacity of indecomposable finite-state channels with feedback. *Proceedings of the Allerton Conf Commun Control Comput*, pp. 1045–1052, September 2008.
- [9] A. Diwadkar and U. Vaidya. Limitations for nonlinear observation over erasure channel. *IEEE Transactions on Automatic Control*, 58: 454 – 459, 2013.
- [10] R. G. Gallager. A simple derivation of the coding theorem and some applications. *IEEE Trans Information Theory*, 11:3–18, 1965.
- [11] A. El Gamal and Y. H. Kim. *Network Information Theory*. Cambridge University Press, U.K., 2012.
- [12] H. Gish and J. N. Pierce. Asymptotically efficient quantization. *IEEE Transactions on Information Theory*, 14:676–683, September 1968.
- [13] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer, Berlin, 2000.
- [14] R. M. Gray. Information rates of autoregressive processes. *IEEE Transactions on Information Theory*, 16:412–421, 1970.
- [15] R. M. Gray. *Entropy and Information Theory*. Springer Verlag, New York, 1990.
- [16] R. M. Gray and T. Hashimoto. A note on rate-distortion functions for nonstationary Gaussian autoregressive processes. *IEEE Transactions on Information Theory*, 54:1319–1322, March 2008.
- [17] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44:2325–2383, October 1998.
- [18] T. Hashimoto and S. Arimoto. On the rate-distortion function for the nonstationary Gaussian autoregressive process. *IEEE Transactions on Information Theory*, 26:478–480, 1980.
- [19] A. Katok. Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. *Inst. Hautes Études Sci. Publ. Math.*, (51):137–173, 1980.
- [20] C. Kawan. State estimation under communication constraints. *arXiv preprint arXiv:1605.03210*, 2016.

- [21] F. Ledrappier and L.-S. Young. Entropy formula for random transformations. *Probability theory and related fields*, 80(2):217–240, 1988.
- [22] D. Liberzon. Stabilization by quantized state or output feedback: A hybrid control approach. In *Proc. IFAC 15th Triennial World Congress*, 2002.
- [23] D. Liberzon and J. P. Hespanha. Stabilization of nonlinear systems with limited information feedback. *IEEE Transactions on Automatic Control*, 50(6):910–915, 2005.
- [24] D. Liberzon and S. Mitra. Entropy and minimal data rates for state estimation and model detection. In *Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control*, pp. 247–256. ACM, 2016.
- [25] T. Linder and S. Yüksel. On optimal zero-delay quantization of vector Markov sources. *IEEE Transactions on Information Theory*, 60:2975–5991, October 2014.
- [26] T. Linder and R. Zamir. Causal coding of stationary sources and individual sequences with high resolution. *IEEE Transactions on Information Theory*, 52:662–680, February 2006.
- [27] N. C. Martins and M. A. Dahleh. Feedback control in the presence of noisy channels: “Bode-like fundamental limitations of performance. *IEEE Transactions on Automatic Control*, 53:1604–1615, August 2008.
- [28] A. S. Matveev. State estimation via limited capacity noisy communication channels. *Mathematics of Control, Signals, and Systems*, 20:1–35, 2008.
- [29] A. S. Matveev and A. Y. Pogromsky. Observation of nonlinear systems via finite capacity channels: Constructive data rate limits. *Automatica*, 70:217–229, 2016.
- [30] A. S. Matveev and A. V. Savkin. *Estimation and Control over Communication Networks*. Birkhäuser, Boston, 2008.
- [31] P. G. Mehta, U. Vaidya, and A. Banaszuk. Markov chains, entropy, and fundamental limitations in nonlinear stabilization. *IEEE Transactions on Automatic Control*, 53(3):784–791, 2008.
- [32] P. Minero, M. Franceschetti, S. Dey, and G. N. Nair. Data rate theorem for stabilization over time-varying feedback channels. *IEEE Transactions on Automatic Control*, 54(2):243–255, 2009.
- [33] G. N. Nair. A nonstochastic information theory for communication and state estimation. *IEEE Transactions on Automatic Control*, 58(6):1497–1510, 2013.
- [34] H. H. Permuter, T. Weissman, and A. J. Goldsmith. Finite state channels with time-invariant deterministic feedback. *IEEE Transactions on Information Theory*, 55(2):644–662, February 2009.
- [35] A. Yu. Pogromsky and A. S. Matveev. A topological entropy approach for observation via channels with limited data rate. *IFAC Proceedings Volumes*, 44(1):14416–14421, 2011.
- [36] A. Sahai. Coding unstable scalar Markov processes into two streams. In *Proceedings of the IEEE International Symposium on Information Theory*, page 462, 2004.
- [37] A. Sahai and S. Mitter. The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link part I: Scalar systems. *IEEE Transactions on Information Theory*, 52(8):3369–3395, 2006.
- [38] M. P. Schutzenberger. On the quantization of finite-dimensional messages. *Information and Control*, 1:153–158, 1958.

- [39] S. Tatikonda and S. Mitter. The capacity of channels with feedback. *IEEE Transactions on Information Theory*, 55(1):323–349, January 2009.
- [40] D. Teneketzis. On the structure of optimal real-time encoders and decoders in noisy communication. *IEEE Transactions on Information Theory*, 52:4017–4035, September 2006.
- [41] U. Vaidya and N. Elia. Stabilization of nonlinear systems over packet-drop links: Scalar case. *Systems & Control Letters*, 61(9):959–966, 2012.
- [42] U. Vaidya and N. Elia. Limitations for nonlinear stabilization over uncertain channel. *preprint*, 2014.
- [43] J. C. Walrand and P. Varaiya. Optimal causal coding-decoding problems. *IEEE Transactions on Information Theory*, 19:814–820, November 1983.
- [44] H. S. Witsenhausen. On the structure of real-time source coders. *Bell Syst. Tech. J.*, 58:1437–1451, July/August 1979.
- [45] Richard G Wood, Tamás Linder, and Serdar Yüksel. Optimal zero delay coding of markov sources: Stationary and finite memory codes. *arXiv preprint arXiv:1606.09135*, 2016.
- [46] S. Yu and P. G. Mehta. Bode-like fundamental performance limitations in control of nonlinear systems. *IEEE Transactions on Automatic Control*, 55(6):1390–1405, 2010.
- [47] S. Yüksel. Stochastic stabilization of noisy linear systems with fixed-rate limited feedback. *IEEE Transactions on Automatic Control*, 55:2847–2853, December 2010.
- [48] S. Yüksel. On optimal causal coding of partially observed Markov sources in single and multi-terminal settings. *IEEE Transactions on Information Theory*, 59:424–437, January 2013.
- [49] S. Yüksel. Stationary and ergodic properties of stochastic non-linear systems controlled over communication channels. *SIAM Journal on Control and Optimization*, pp. 2844–2871, 2016.
- [50] S. Yüksel and T. Başar. *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. Birkhäuser, New York, NY, 2013.
- [51] S. Yüksel and S. P. Meyn. Random-time, state-dependent stochastic drift for Markov chains and application to stochastic stabilization over erasure channels. *IEEE Transactions on Automatic Control*, 58:47–59, January 2013.
- [52] P. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28:139–148, March 1982.
- [53] H. Zang and P. A. Iglesias. Nonlinear extension of Bode’s integral based on an information-theoretic interpretation. *Systems & control letters*, 50(1):11–19, 2003.
- [54] J. J. Zhu. Two notes on measure-theoretic entropy of random dynamical systems. *Acta Mathematica Sinica*, 25(6):961–970, 2009.